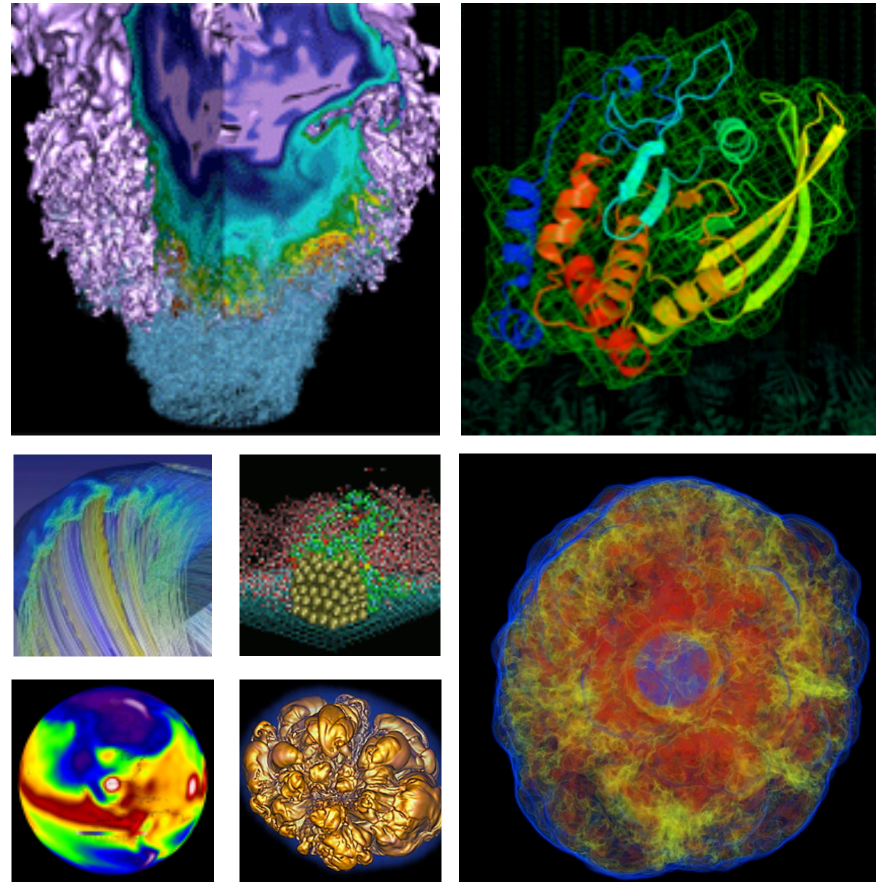
2014 NERSC Workload Analysis



Brian Austin, Wahid Bhimji, Tina Butler,
Jack Deslippe, Scott French, Richard Gerber
Douglas Jacobsen, Nicholas Wright,
Zhengji Zhao

November 5, 2015

NERSC

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Understanding the NERSC workload is key to procuring productive, high performing systems for science.

- **Conducted workload analysis to understand application requirements and guide future system procurements.**
- **Important for understanding efforts needed to transition workload to future architectures.**
- **Analyzed the workload by:**
  - Science area
  - Application code
  - Algorithm
  - Job size
  - Thread usage
  - Memory usage
  - Library usage
  - I/O usage

Workload analysis aims to understand how users exercise the available computational resources.

NERSC engages in other activities to complement the workload analysis.

- **Requirement reviews ascertain the future needs of users.**
- **Benchmarking and performance analysis reveals performance characteristics and sensitivities of individual applications.**
- **Work*flow* analysis describes the operational and data dependencies of a single project. (The work*load* is a cross-section of many simultaneous workflows.)**

Requirements for future procurements are obtained by combining *all* these sources of information. A retrospective workload analysis reflects current (not future) hardware and software resource utilization.

Methods

Data collected in this presentation came from a variety of sources.

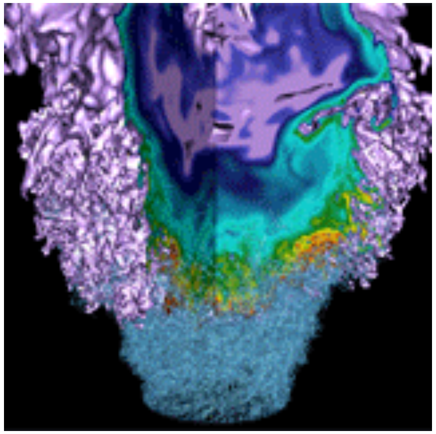
- System accounting logs
- NIM database
- ALPS command line capture log
- Automatic Library Tracking Database (ALTD)
- Resource Utilization Report (RUR)
- Lustre Monitoring Tool (LMT)

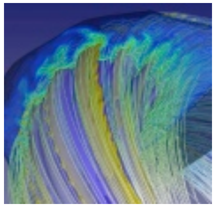
Current (and imminent !) NERSC systems.

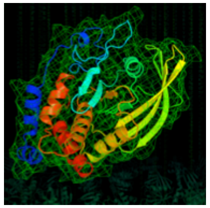
	Hopper Cray XE6 (2011)	Edison Cray XC30 (2013)	Cori Cray XC40 (2016)
			
Interconnect	6384 nodes Cray Gemini (3D Torus)	5576 nodes Cray Aries (Dragonfly)	9300 KNL nodes plus 1624 Haswell nodes Cray Aries (Dragonfly)
Processor	Two 12-core AMD Magny Cours (2.1 GHz)	Two 12-core Intel Ivy-Bridge (2.4 GHz)	One 64+ core Intel Knight's Landing (GHz TBD)
Memory	32 GB/node; 54 GB/s	64 GB/node; 102 GB/s	96 GB DDR4/node; 90 GB/s 16 GB HBM; >400 GB/s
Scratch Filesystem	2.0 PB; 70 GB/s	7.5 PB; 168 GB/s	28.5 PB; >700 GB/s Burst Buffer: 1.5 PB; 1.5 TB/s
Sustained System Performance*	144 Tflop/ s	293 Tflop/s	>10 x Hopper

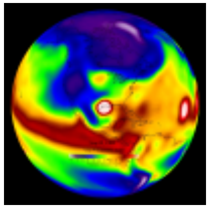
* https://www.nersc.gov/users/computational-systems/edison/performance-and-optimization/performance-comparison-between-edison-and-hopper/

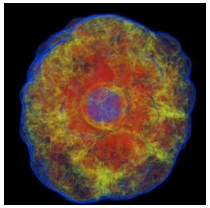
Workload Diversity

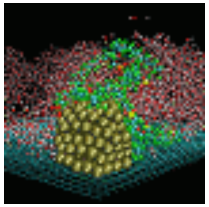










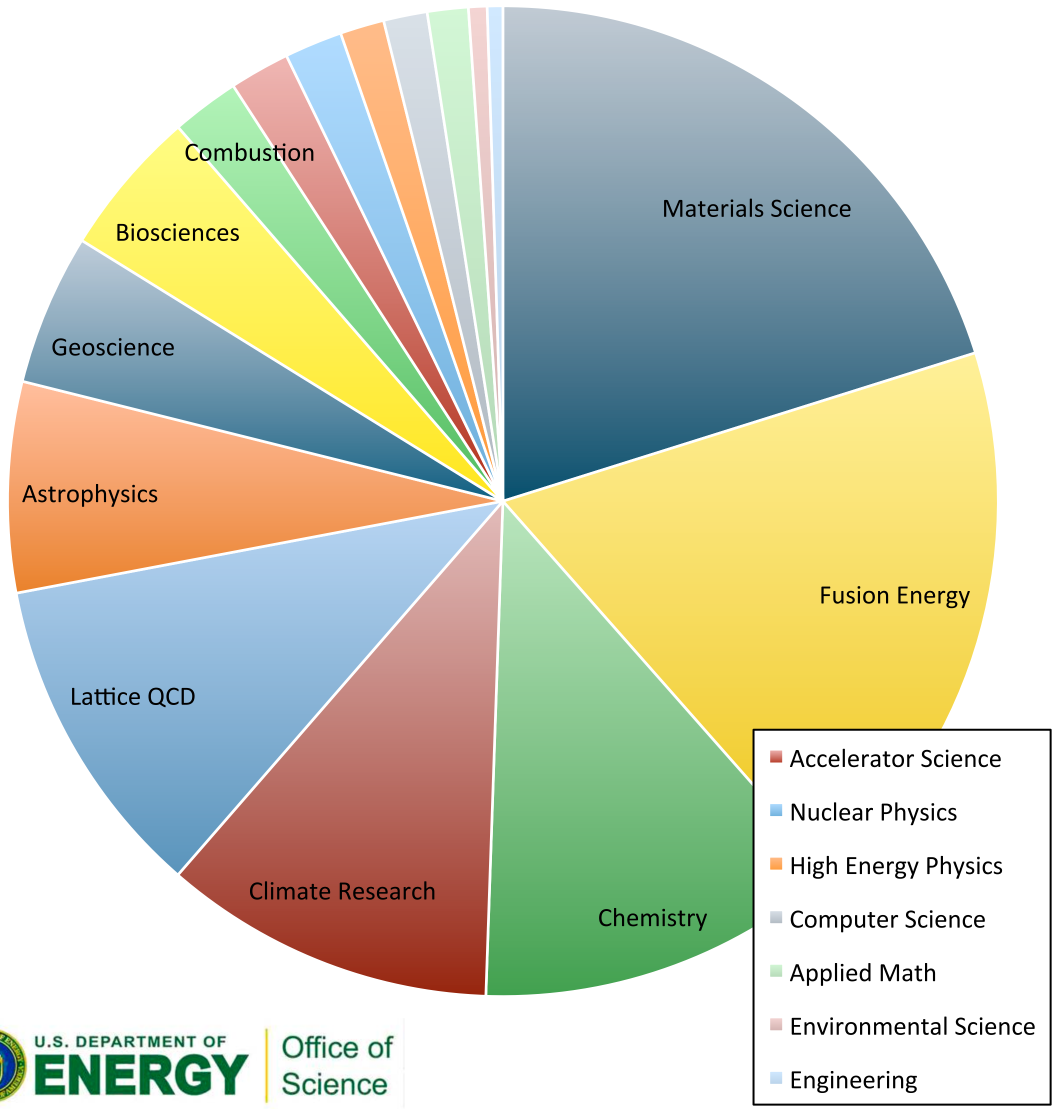


Workload diversity questions:

- **Which science domains and algorithms are represented in the applications at NERSC?**
- **What codes, libraries and languages are most important to NERSC users?**

NERSC serves a broad range of science disciplines for the DOE Office of Science

Workload distribution by 2014 allocation



- **Over 5950 users**
- **Nearly 850 projects**

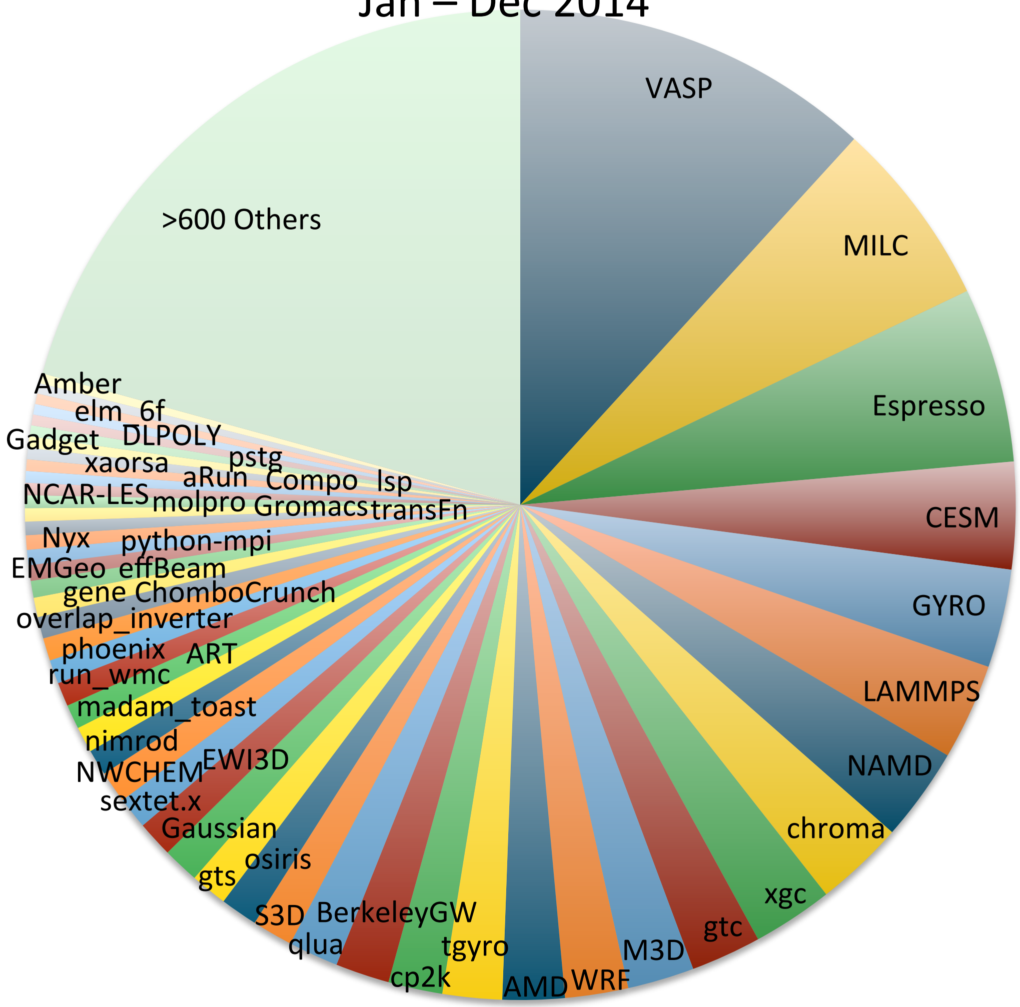
Top 5 Science Categories by allocation (2014)	
Materials Science	20%
Fusion Energy	18%
Chemistry	12%
Climate Research	11%
Lattice QCD	11%

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

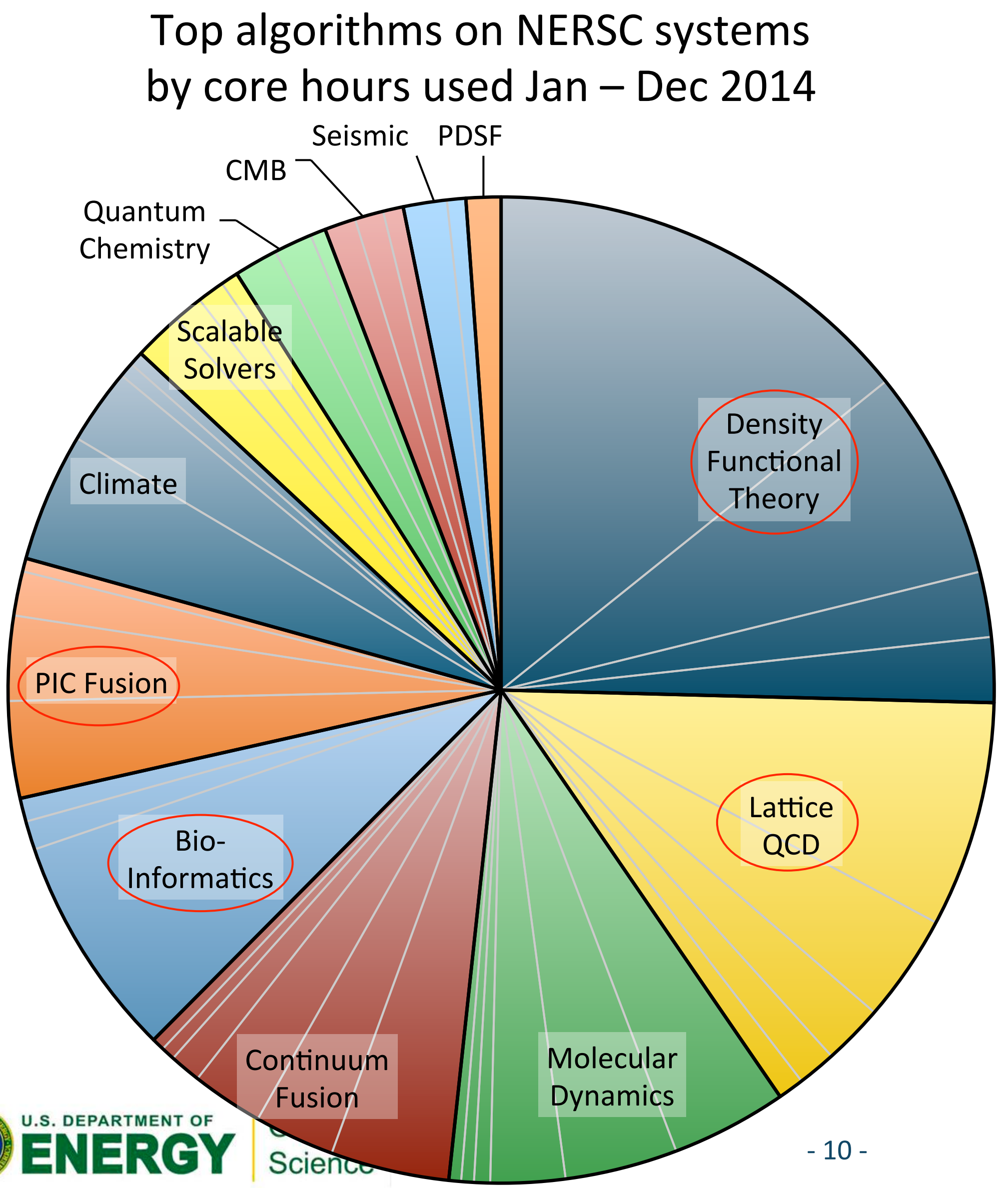
Over 650 applications run on NERSC resources

Top Application codes on Hopper and Edison by hours used.
Jan – Dec 2014



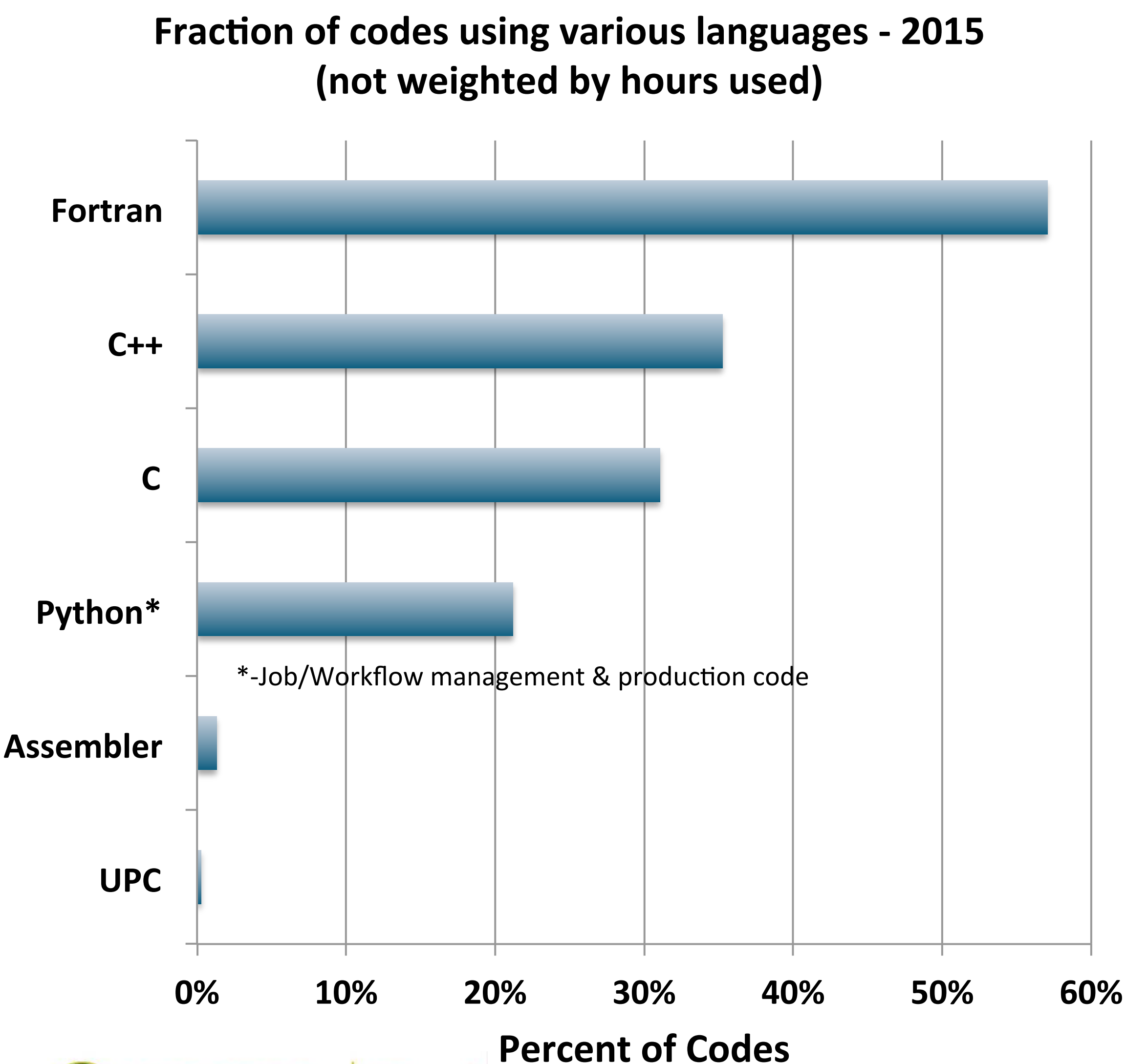
- **13 codes make up 50% of workload**
- **25 codes make up 66% of workload**
- **50 codes make up 80% of workload**
- **Remaining codes (over 600) make up 20% of workload.**

Many codes implement similar algorithms.



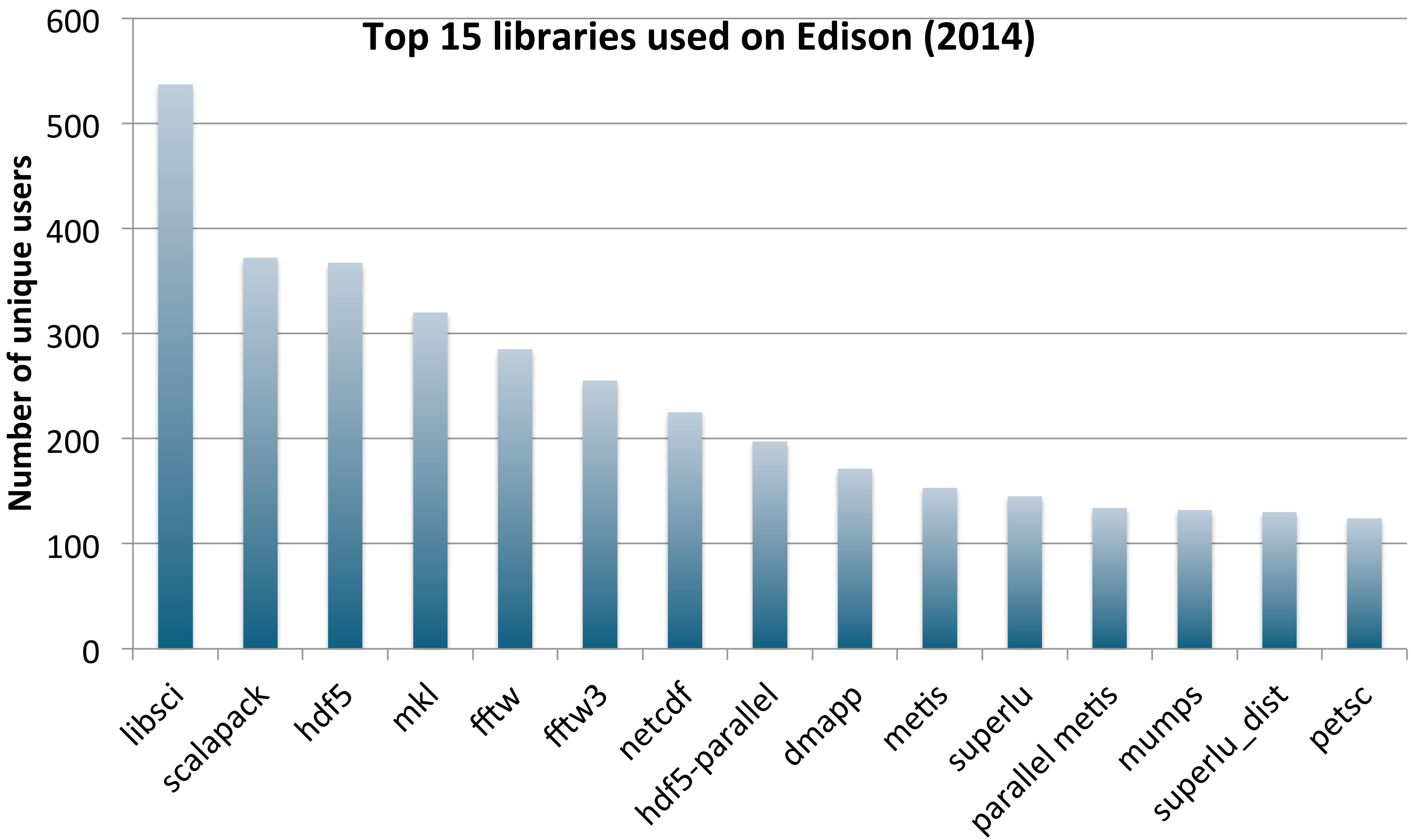
- **Regrouped top codes by similar algorithms.**
- **A small number of benchmarks can represent a large fraction of the workload.**
- **Includes Genepool and PDSF systems.**
  - Carver was similar in size to PSDF, but had a diverse workload.

Languages Used at NERSC



- **Based on user surveys.**
- **Fortran would be even more important if codes were weighted by hours used.**
  – Fortran is the primary language for 23 of the 36 top codes.
- **Total exceeds 100% because some codes use multiple languages.**

NERSC's broad workload relies on optimized libraries to maximize performance.

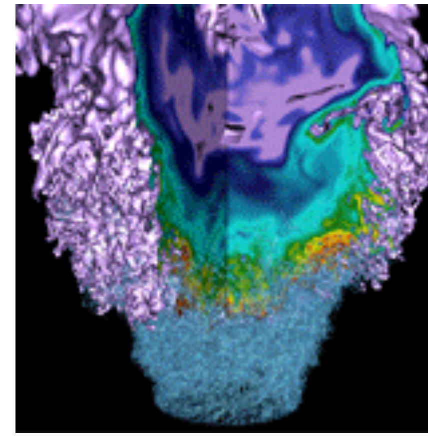


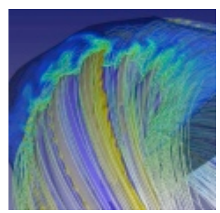
NERSC enables a prodigious volume of scientific research.

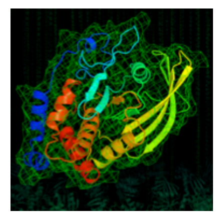
NERSC

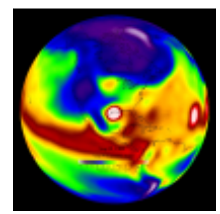
- **Over 1800 publications during 2014**

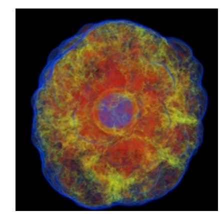
Concurrency

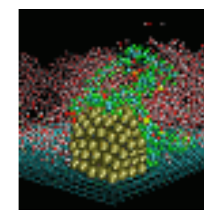








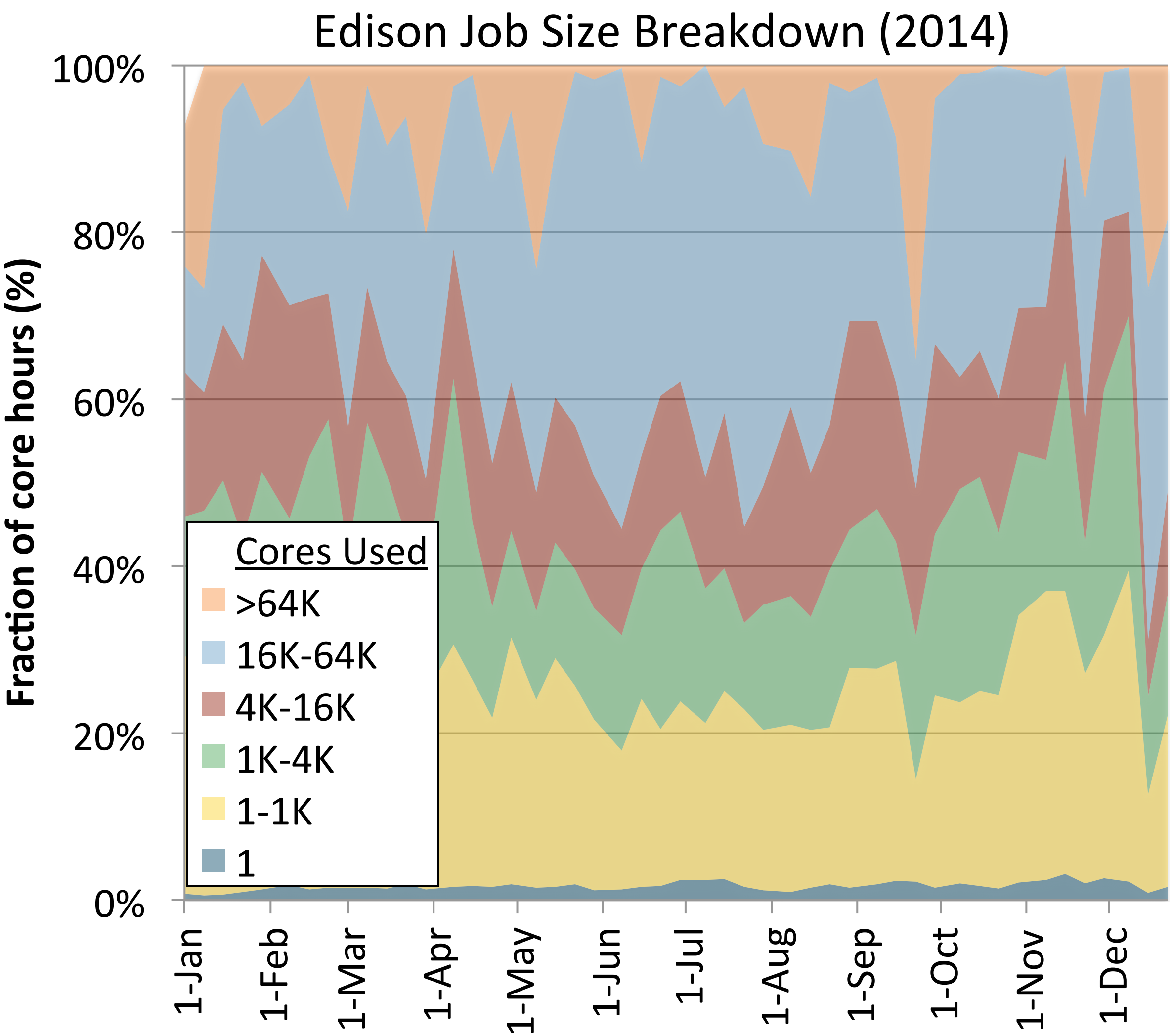




Parallelism and Concurrency

- **What are common job sizes at NERSC?**
- **How are users expressing parallelism in their codes?**
- **Users will likely need threads to take full advantage of many-core architectures like Cori. How much is OpenMP used now?**

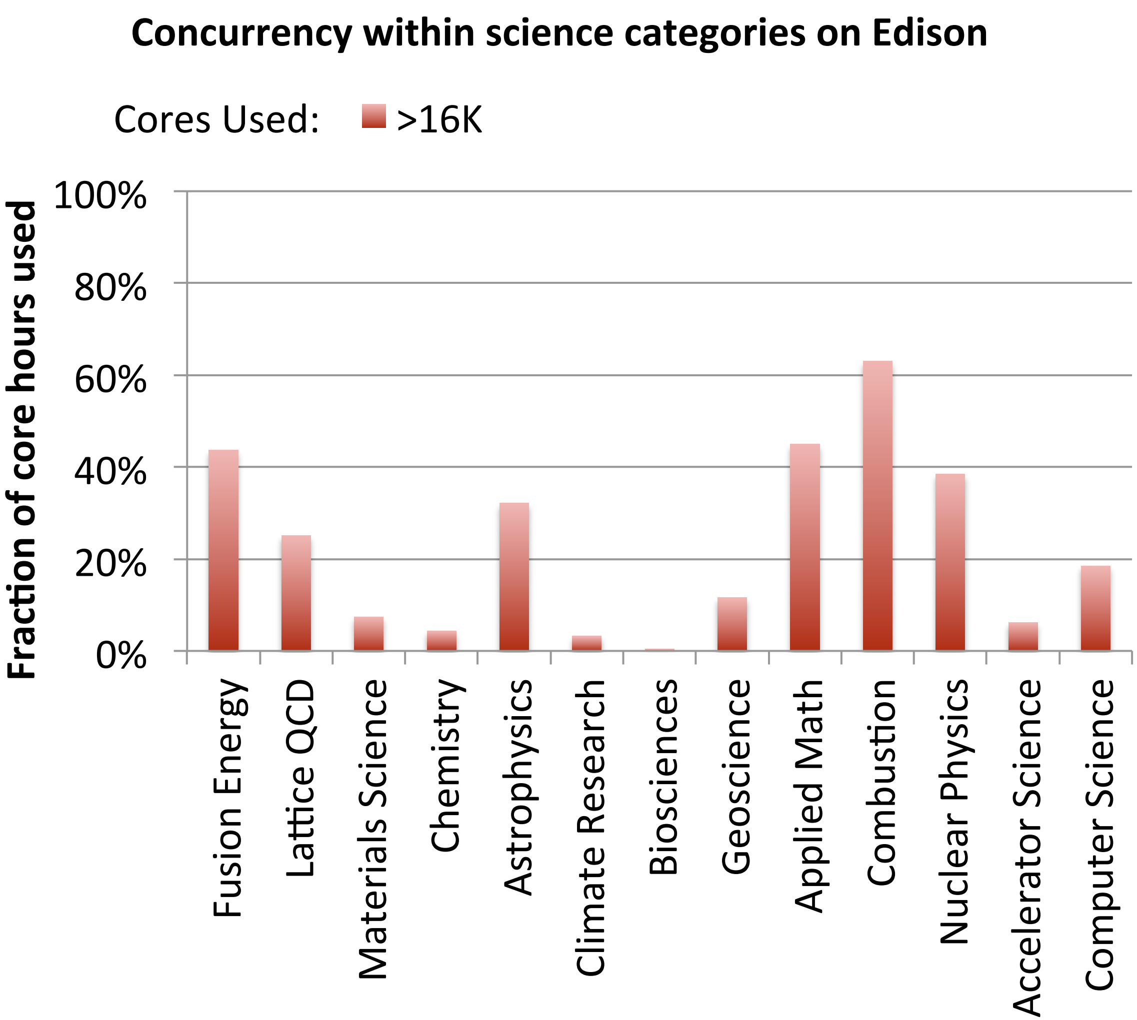
High concurrency jobs are a significant fraction of the NERSC workload.



- **37% of Edison hours use more than 16 K cores.**
- **4% of Edison hours use more than 2/3 of its cores.**

Cores	Core Hours
64 K – 100%	7%
16 K – 64 K	31%
4 K – 16 K	17%
1 K – 4 K	18%
1 – 1 K	25%
1	2%

High concurrency jobs are used in all science domains.

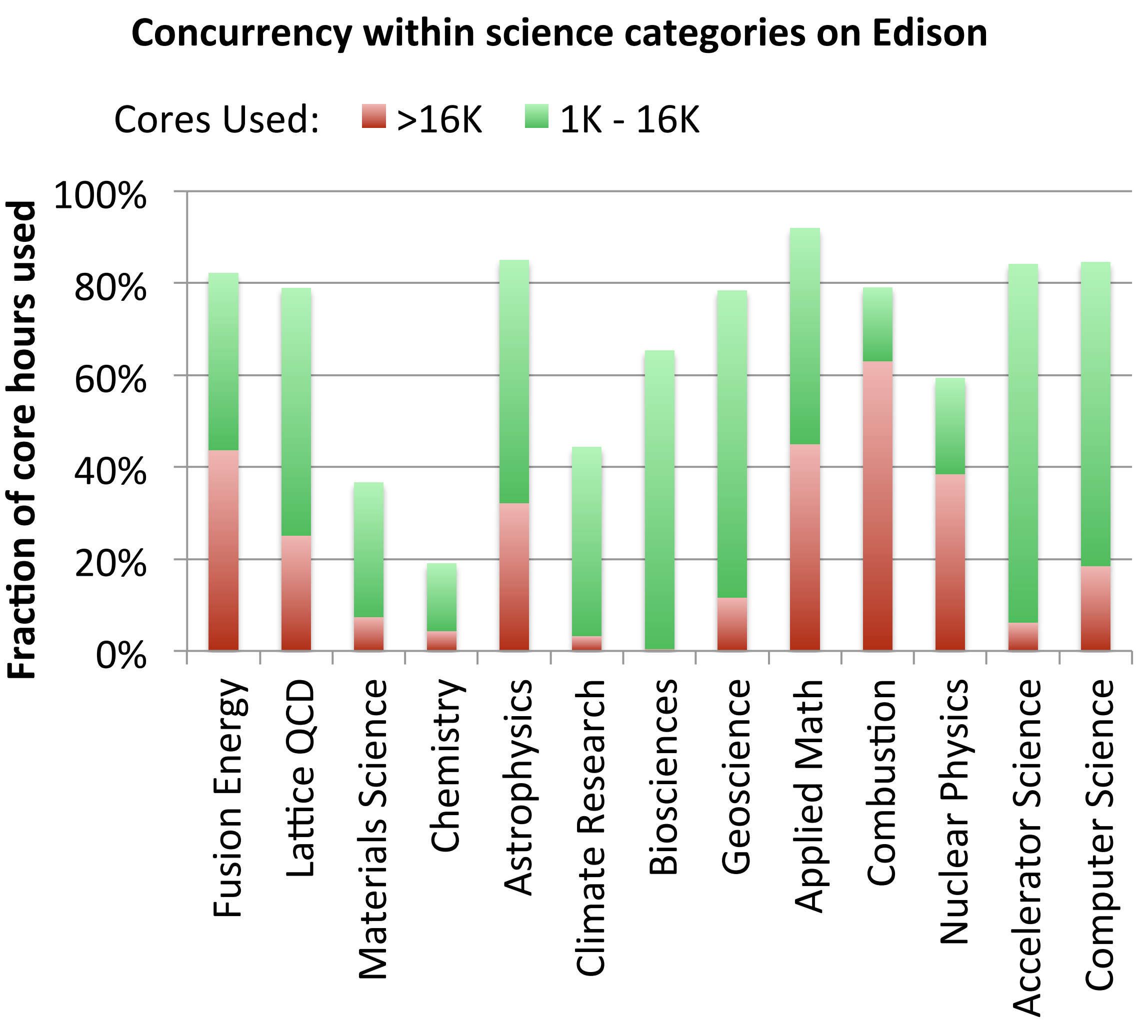


- **Some fraction of every domain's workload runs with more than 16K cores.**

U.S. DEPARTMENT OF ENERGY | Office of Science

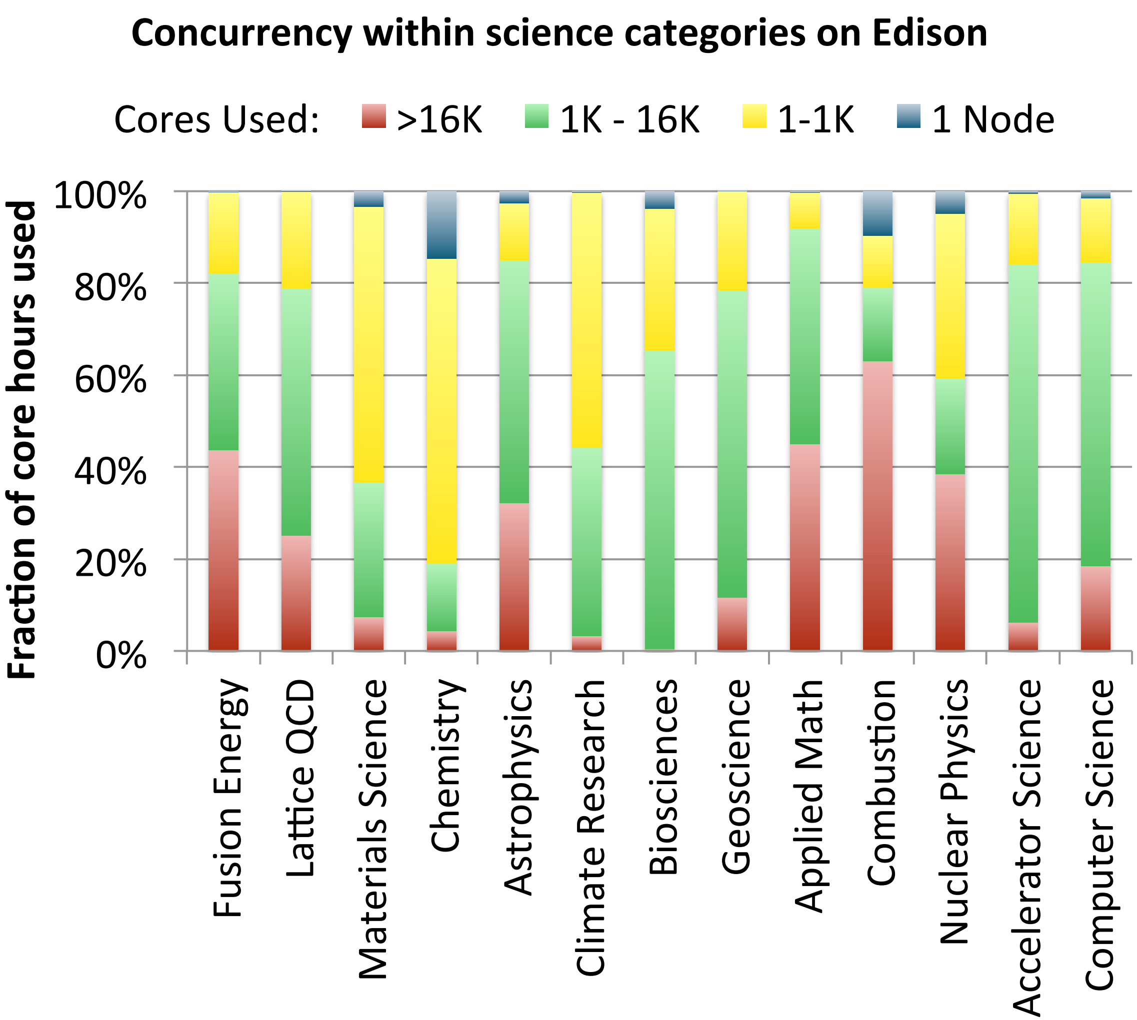
BERKELEY LAB

High concurrency jobs are used in all science domains.



- Some fraction of every domain's workload runs with more than 16K cores.
- In almost all domains, more than half the workload uses more than 1K cores.

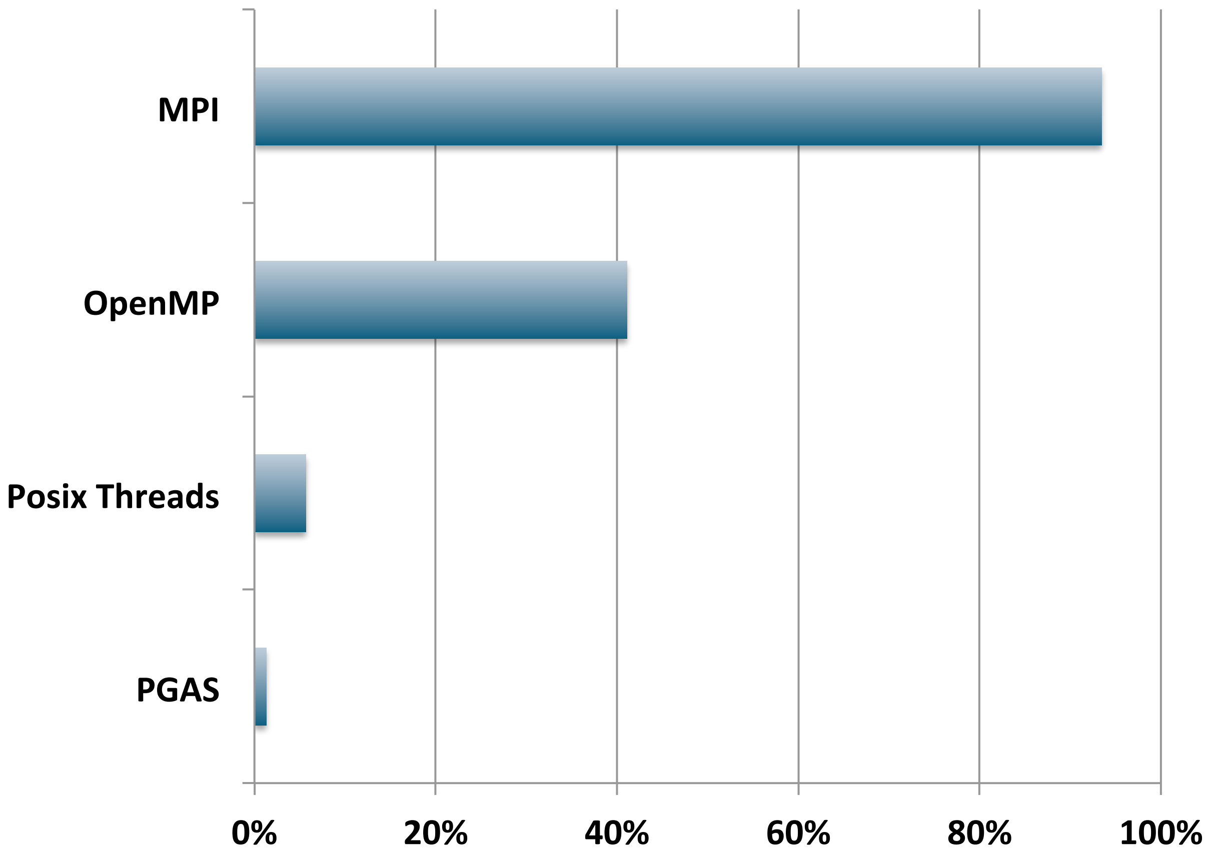
High concurrency jobs are used in all science domains.



- **Some fraction of every domain's workload runs with more than 16K cores.**
- **In almost all domains, more than half the workload uses more than 1K cores.**
- **Does not include the Genepool or PDSF clusters.**
  - Combined, these are 7% of the workload.

Nearly all projects rely on MPI for distributed memory parallel programming.

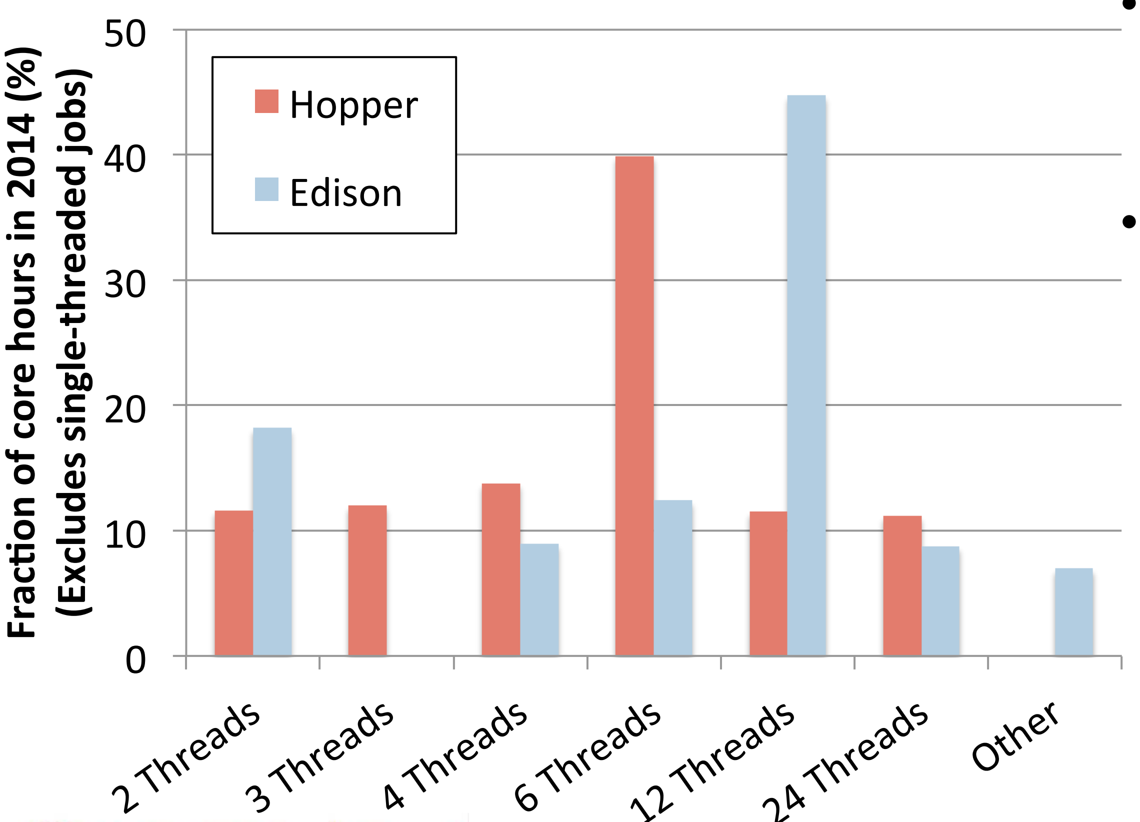
Fraction of codes using various parallel programming models.



- Based on user survey of codes used. Not weighted by core hours.
- Total exceeds 100% because some codes use multiple languages.
- 40% of projects *report* using OpenMP.

NERSC users are embracing threads.

	Hopper	Edison
Fraction of hours using OpenMP	14%	21%

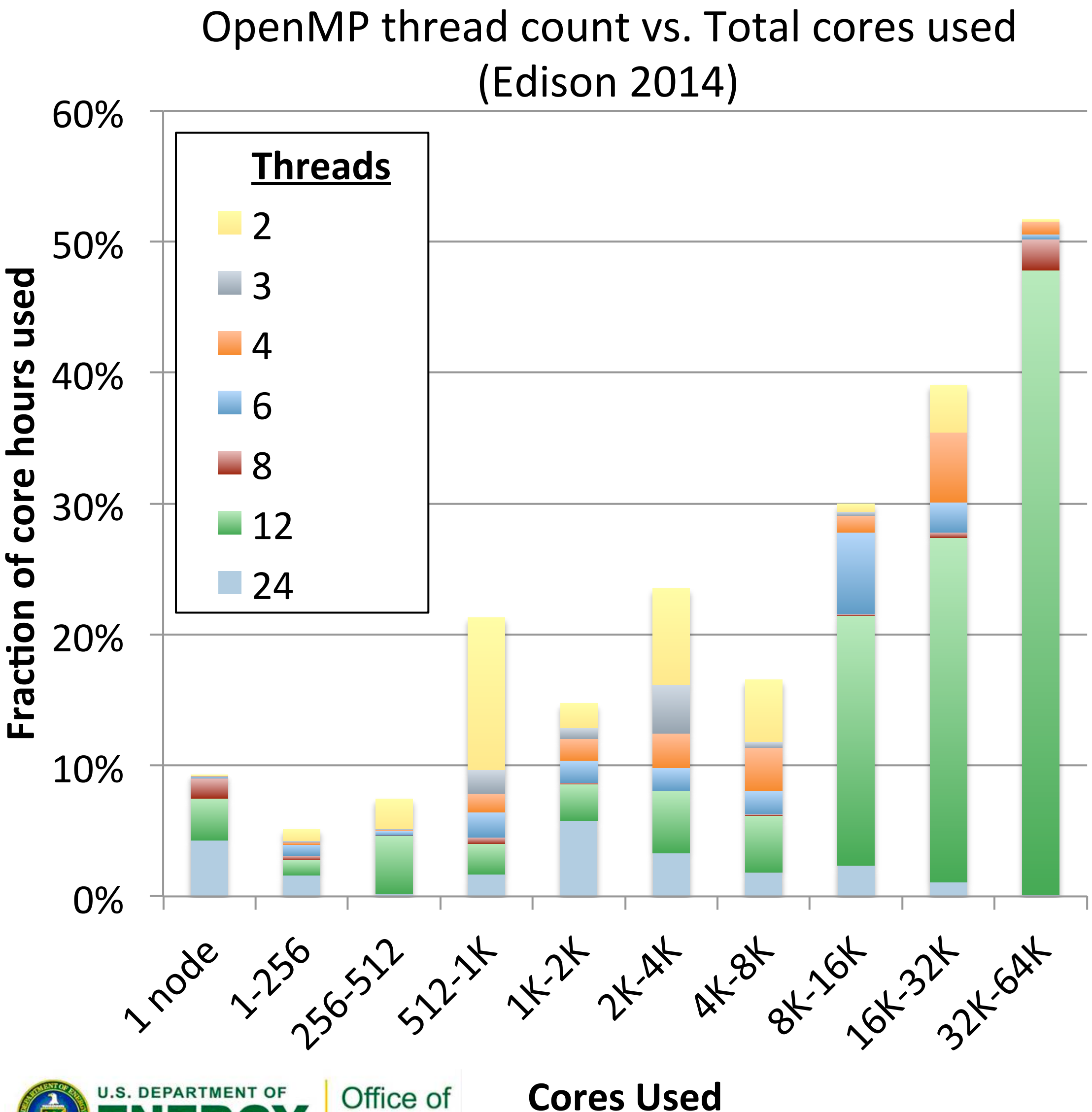


- **Currently nearly 20% of hours are consumed using multiple OpenMP threads.**
- **Thread concurrency has increased over generations of systems.**
- **On both systems, the dominant thread concurrency matches the NUMA domain.**

  Hopper: 6 cores per NUMA domain

  Edison: 12 cores per NUMA domain

High concurrency jobs use more threads.



- **Thread utilization increases with node count.**
  - More than half of the core hours using 2/3 of Edison are threaded. (not shown)
- **Thread concurrency increases with node count.**
  - Jobs with 12 threads per process is dominate at higher concurrency.
- **OpenMP use increases at large scales where MPI scaling inefficiencies outweigh (on-node) OpenMP inefficiencies.**

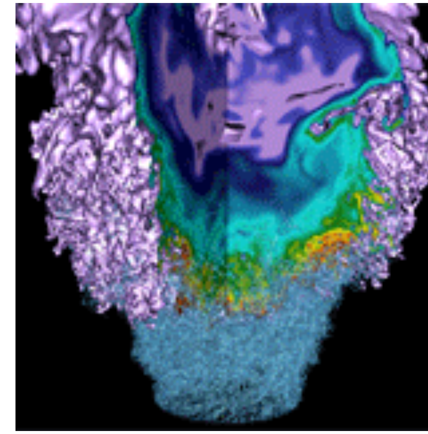
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

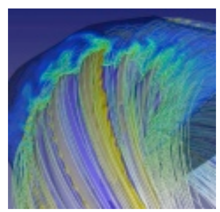
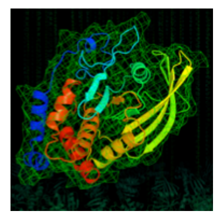
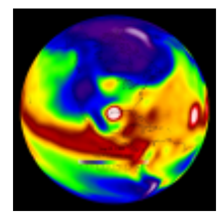
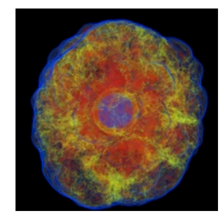
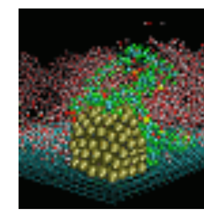
Summary

- **Users need to run single-node jobs, full-system jobs, and everything in between.**
  - 37% of the Edison workload use more than 16k cores
  - 75% uses more than 1024 cores.
- **MPI is (still) the predominant form of parallelism in user codes.**
- **About 20% of the workload uses threads.**
  - OpenMP adoption has increased over system generations.
  - Thread utilization increases with node count.
  - Thread concurrency seems to match NUMA domain size.

Memory utilization



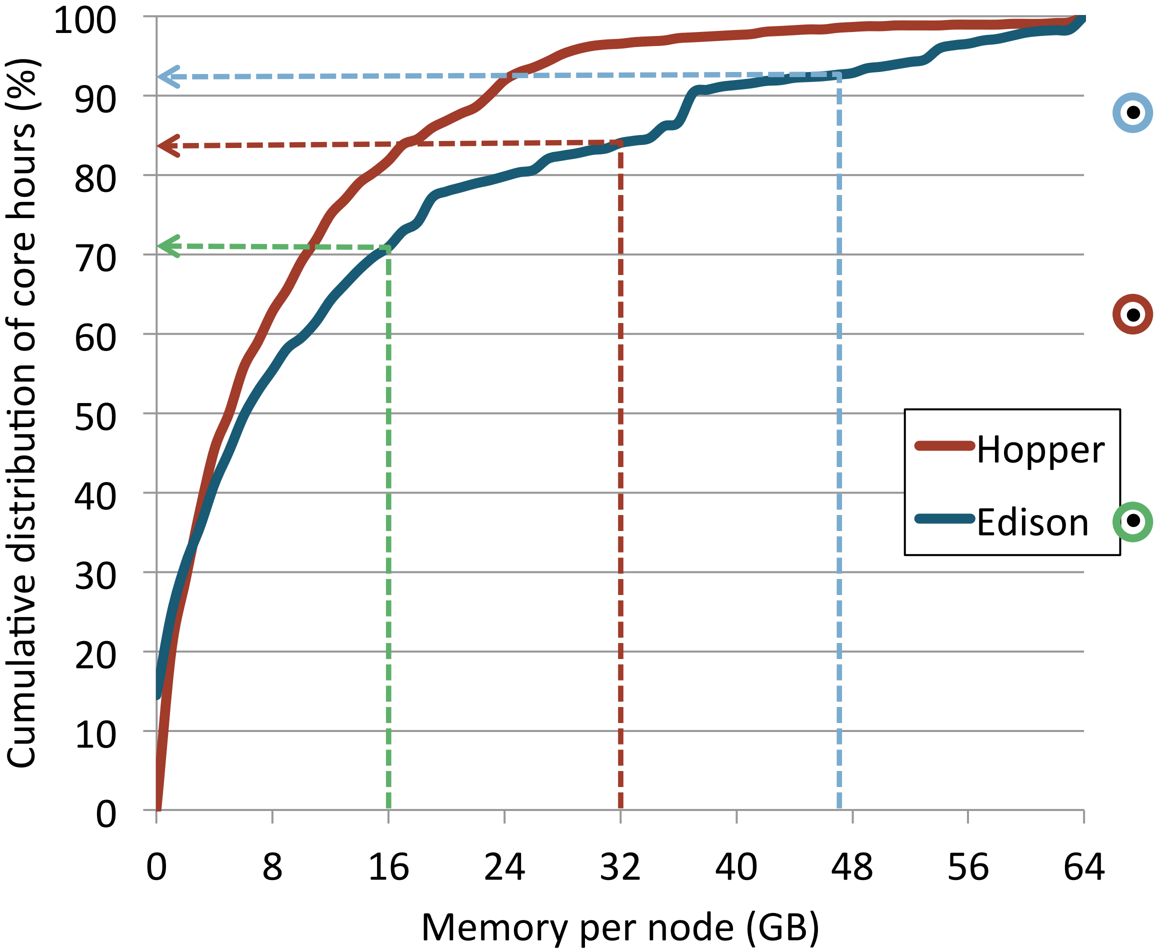
NERSC

Memory utilization

- How much memory is being used per node? Per MPI rank?
- Edison has twice as much memory per node as Hopper. How often is it used?
- What fraction of the NERSC workload will fit into Cori's HBM without modification?
- Limited memory (and HBM) capacity was a potential motivator for thread adoption. Is this reflected by current OpenMP use?

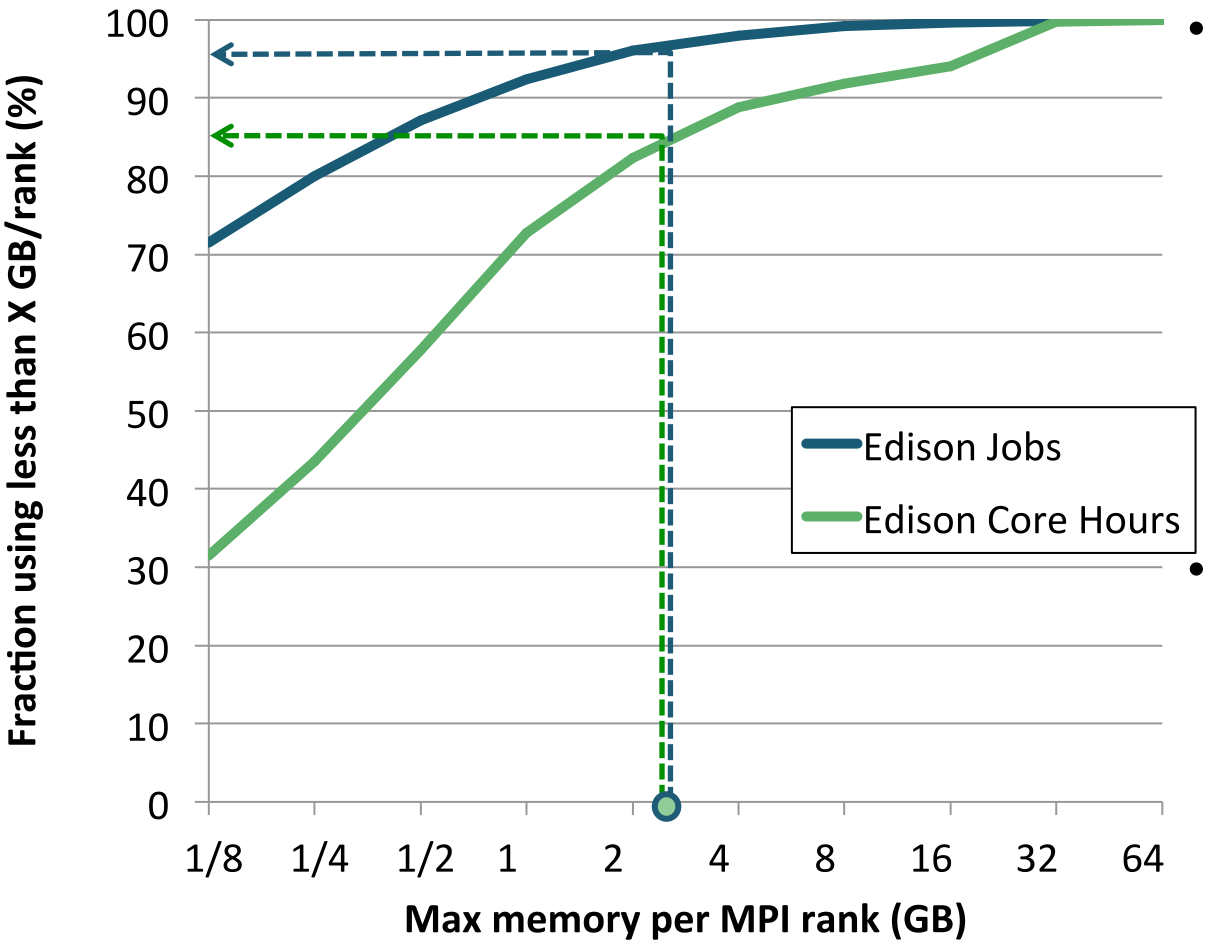
Users are taking advantage of Edison's increased memory per node.



- Hopper has 32 GB nodes, Edison has 64 GB nodes
- 8% of Edison workload uses more than 80% of available memory per node.
- 16% of the Edison workload would not run on Hopper's 32 GB nodes.*
- 71% of Edison workload will fit into Cori's fast memory (16 GB).

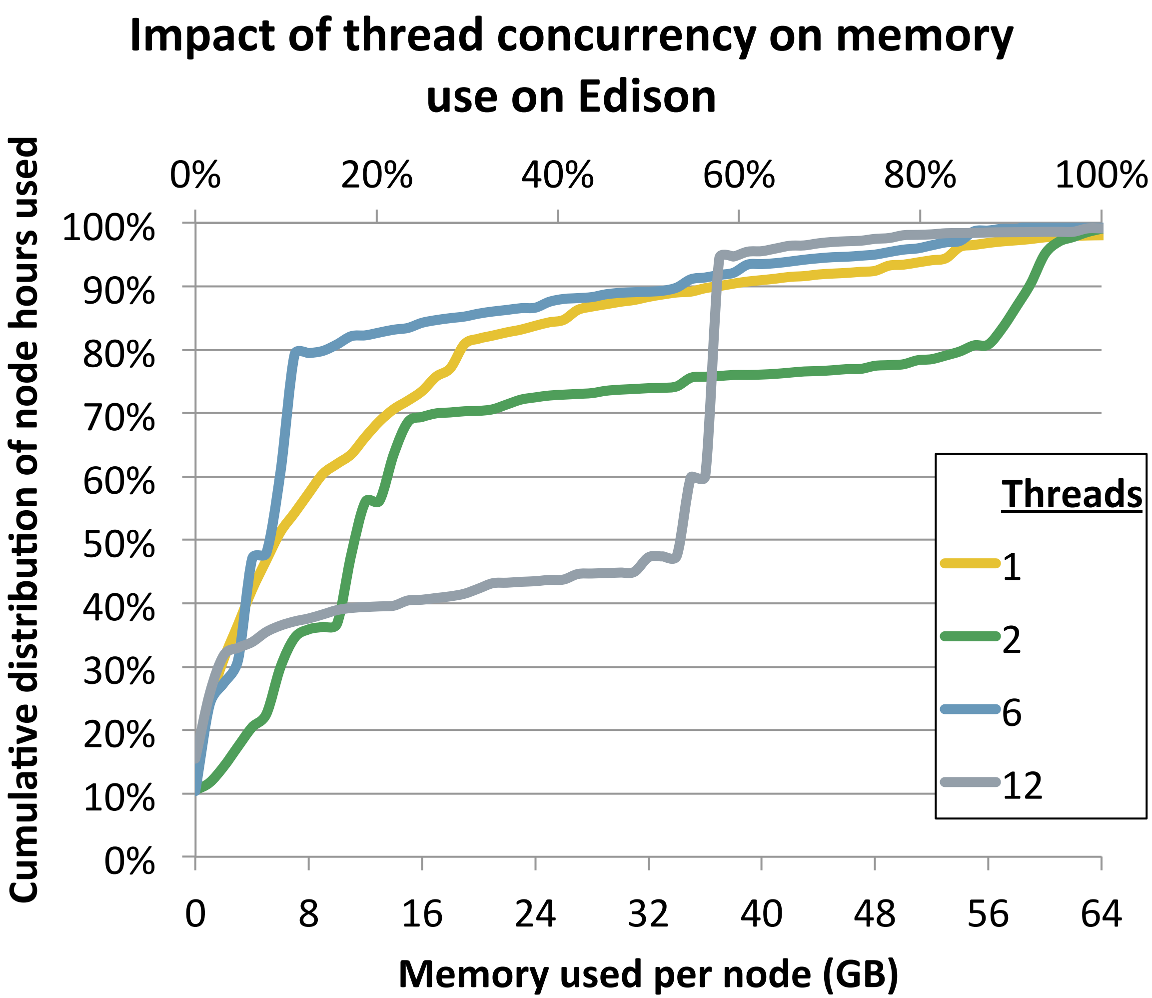
*Assuming MPI+X concurrency does not change.

A modest fraction (10%) of the Edison workload uses more than 4 GB per MPI rank.



- **Most Edison users are not constrained by memory capacity.**
  - 15% of Edison hours use more than 2.6 GB / rank.
  - Of this 15%, four threaded codes make up 60%.
  - Much of the remaining 40% is sequential code
- **Many users run a handful of large memory jobs.**

OpenMP adoption does not seem to be driven by limited memory capacity.

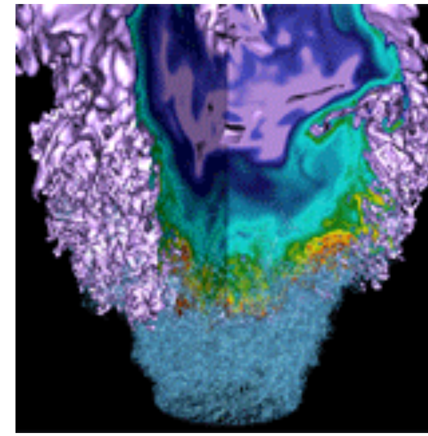


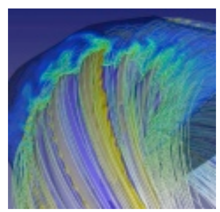
- Only a small fraction (<5%) of multi-threaded jobs use more than 80% of available memory.
- Most (>95%) multi-threaded jobs have sufficient memory to accommodate an additional MPI rank per node.
- No simple relationship between thread concurrency and memory use.

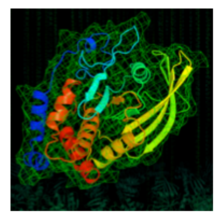
Memory capacity summary

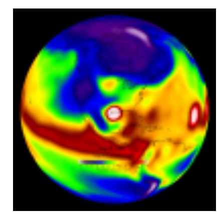
- About 1/6th Edison's workload could not fit into Hopper's 32 GB nodes.
- About half of the Edison workload will have no problems running exclusively in Cori's HBM (assuming no changes).
- OpenMP adoption does not seem to be driven by limited memory capacity.

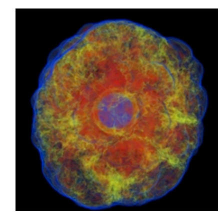
Storage and I/O

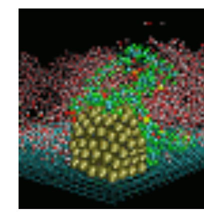








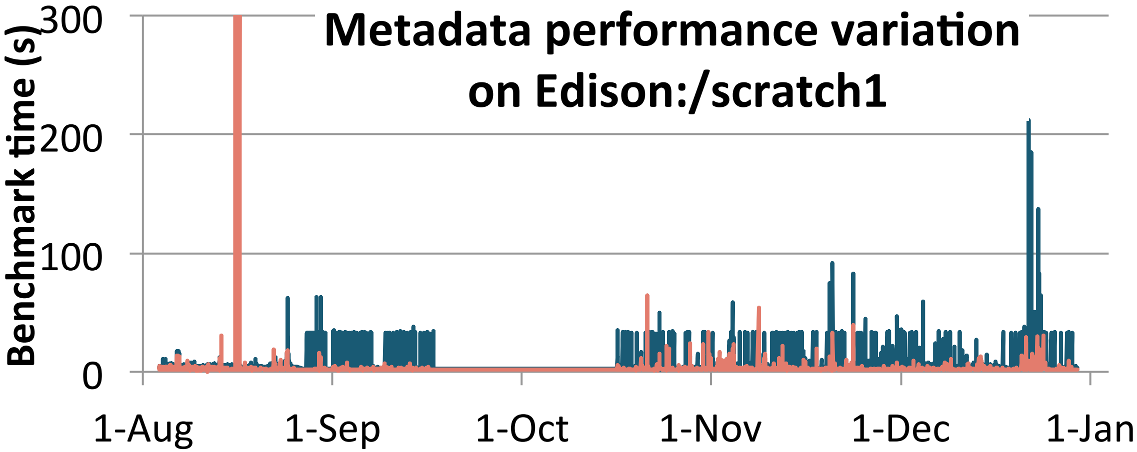


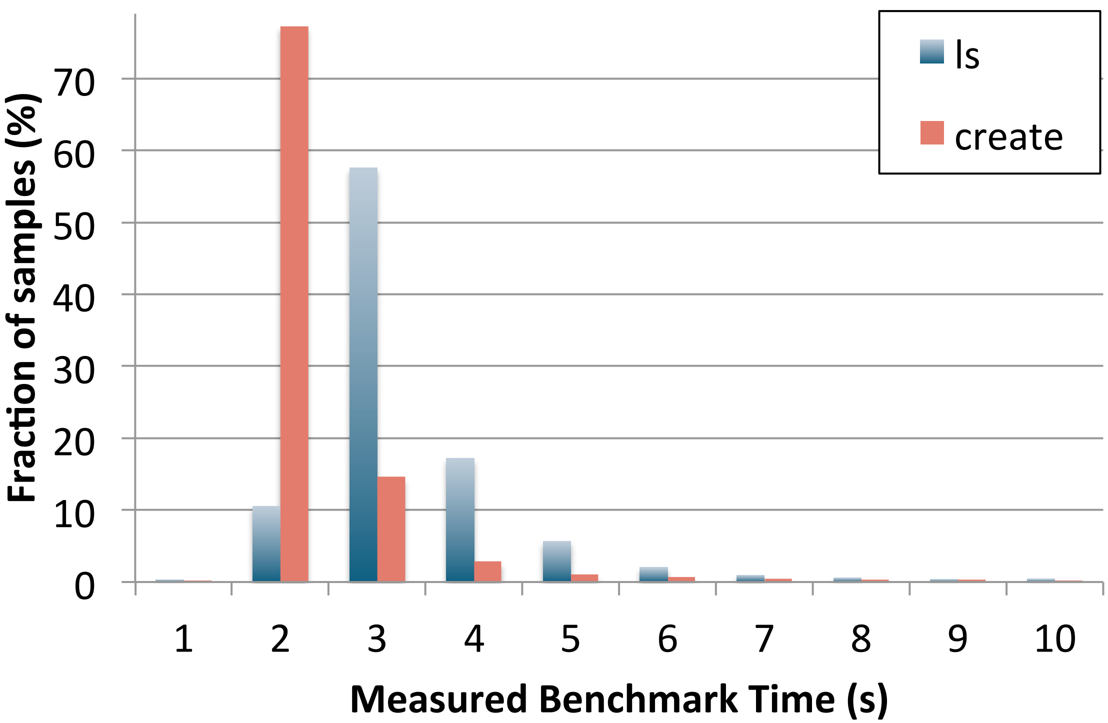


Storage and I/O questions

- **What are the biggest I/O issues effecting users?**
- **What are the read and write volumes of filesystem activity?**
- **How much of the I/O load is due to checkpointing?**
- **How quickly are NERSC filesystems filling?**
- **What is the distribution of file sizes?**

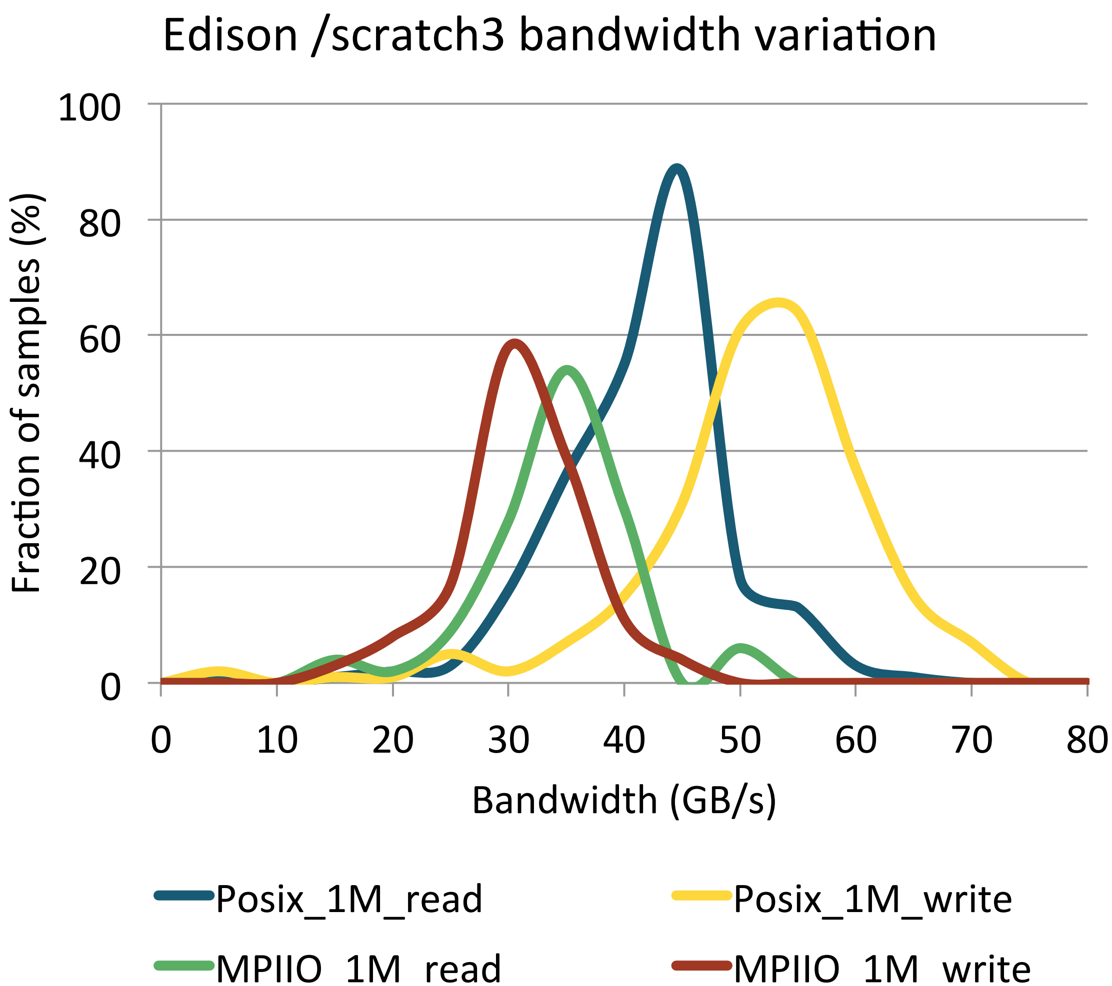
More reliable metadata performance would improve application performance.





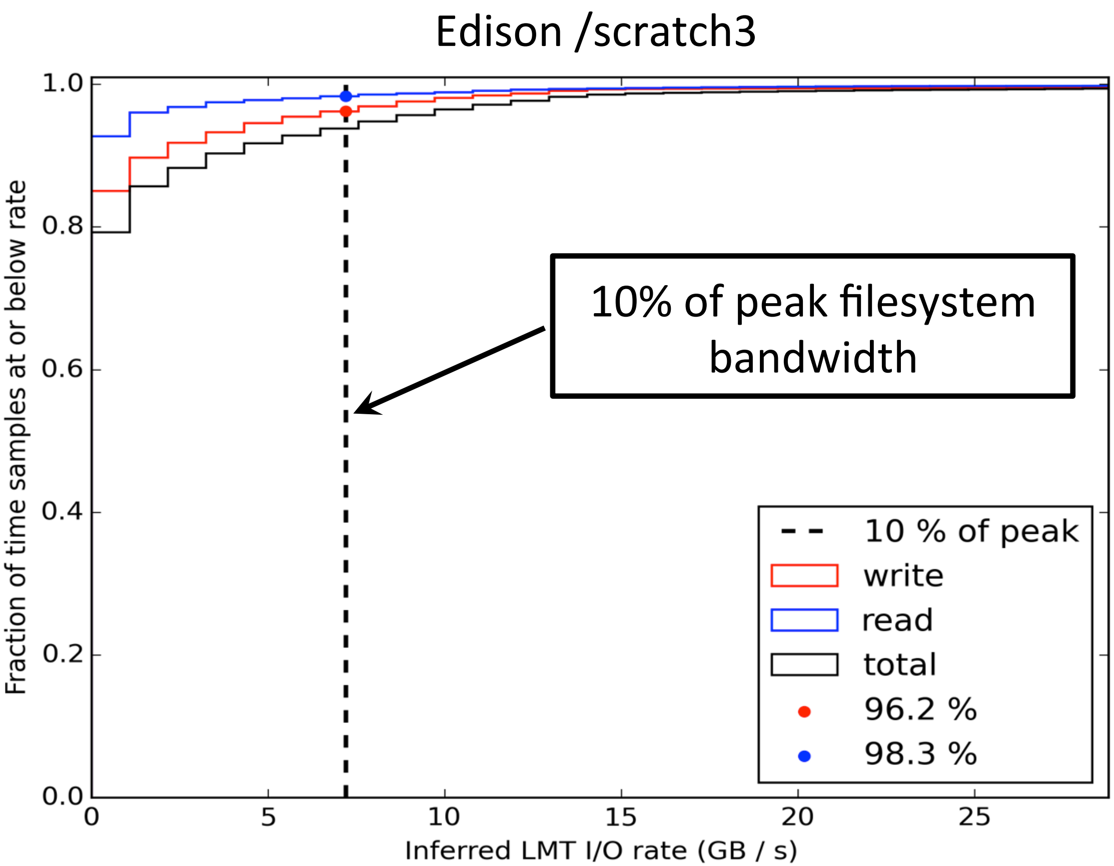
- Cron job times "ls" and file creation every five minutes to test I/O metadata performance on Edison's scratch1 filesystem.
- Benchmarks normally complete in 2 or 3 seconds.
- More than one in five tests are significantly slower.
- Both benchmarks have long tails stretching to 300s.

I/O bandwidth variation degrades quality of service



- **Cron job measures performance of IOR benchmark each week.**
- **I/O benchmarks routinely measure large fractions of peak bandwidth.**
- **"Typical" measurements are 25-40% slower.**
  - 30-50% variation
  - A few runs are much slower.

Users seldom achieve large fractions of peak I/O bandwidth.



- **Lustre Monitoring Tool (LMT) counts total data read/written within 5 second intervals.**
- **Even poorly performing benchmark runs exceed the I/O rates observed in production.***
  - No file system exceeds 10% of peak more than 10% of the time.
  - 99% of /scratch3 samples use less than 20 GB/s (27% of peak).

  *Actual I/O rates may exceed the inferred rates. (Large sampling window)

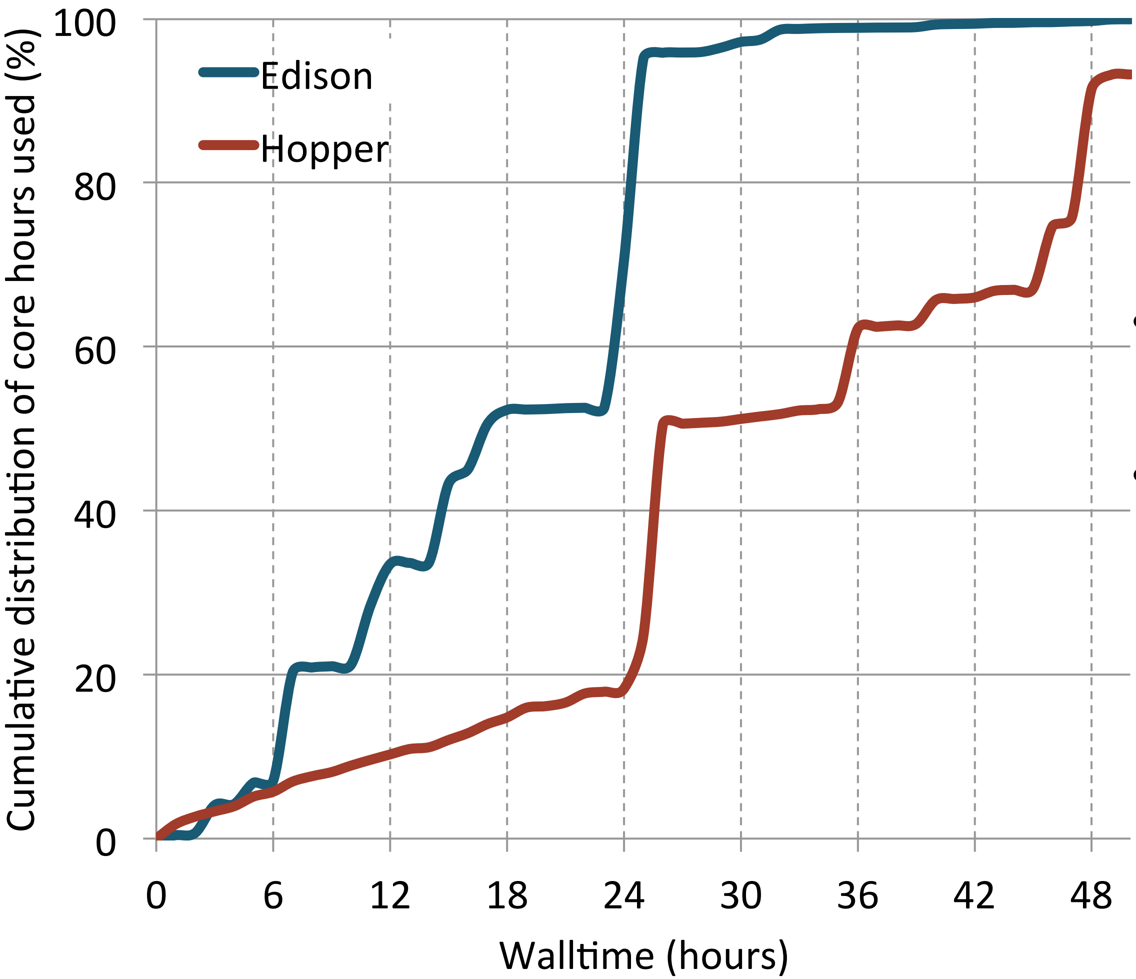
- **Significant fractions of peak are routinely measured.**
  - See benchmark results on previous slide.
  - 63 of 812,000 LMT samples exceed 80% of peak on Edison's /scratch3.

Maximum daily write volume ≈ 2× memory capacity.

- **LMT measurements of data read/written each day, summed over scratch filesystems.**
- **Read/write balance shifts from Hopper to Edison.**
  - Read volume is similar between systems.
  - Edison has 3x write volume.

Average daily scratch I/O volume (TB)		
	Read	Write
Hopper	139.8	105.2
Edison	139.4	303.0

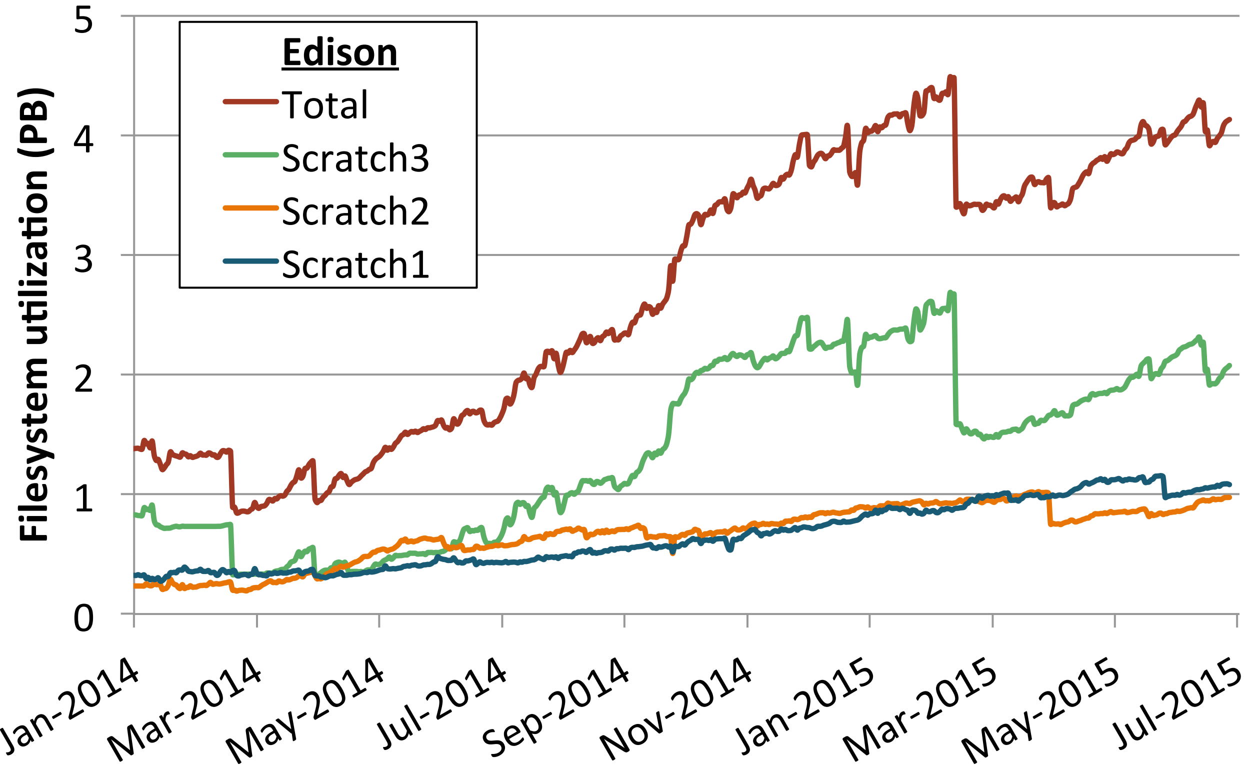
Much of the NERSC workload seems to use checkpoint-restart functionality.



- **A large fraction (70%) of core hours is consumed by jobs that reach the wallclock limit.**
  - Steps in plot correspond to queue limits.
- **Users want longer queues**

  (and shorter wait times)
- **95% of jobs run for less than one hour.**

Edison scratch filesystem overview



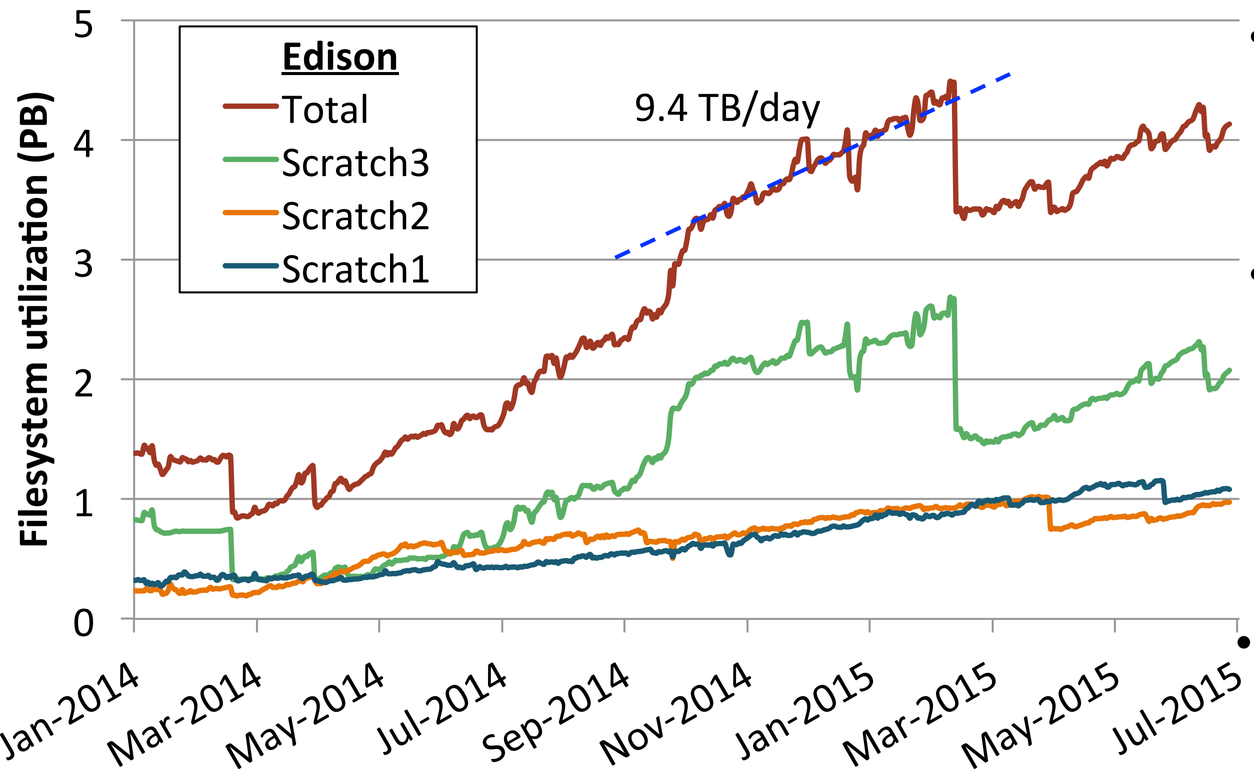
Filesystem	/scratch1	/scratch2	/scratch3	Total
Capacity (TB)	2100	2100	3200	7400
Bandwidth (GB/s)	48	48	72	168

- **Edison has three scratch filesystems.**
- **Users are randomly assigned to either /scratch1 or /scratch2**
  - Performance isolation
  - Improved metadata performance
- **Users with demanding I/O requirements may opt-in to /scratch3.**
  - 1.5x bandwidth
  - 1.5x capacity
  - Default striping increased for better bandwidth.
  - Additional performance isolation

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Edison scratch filesystem utilization increases 10 TB/day.



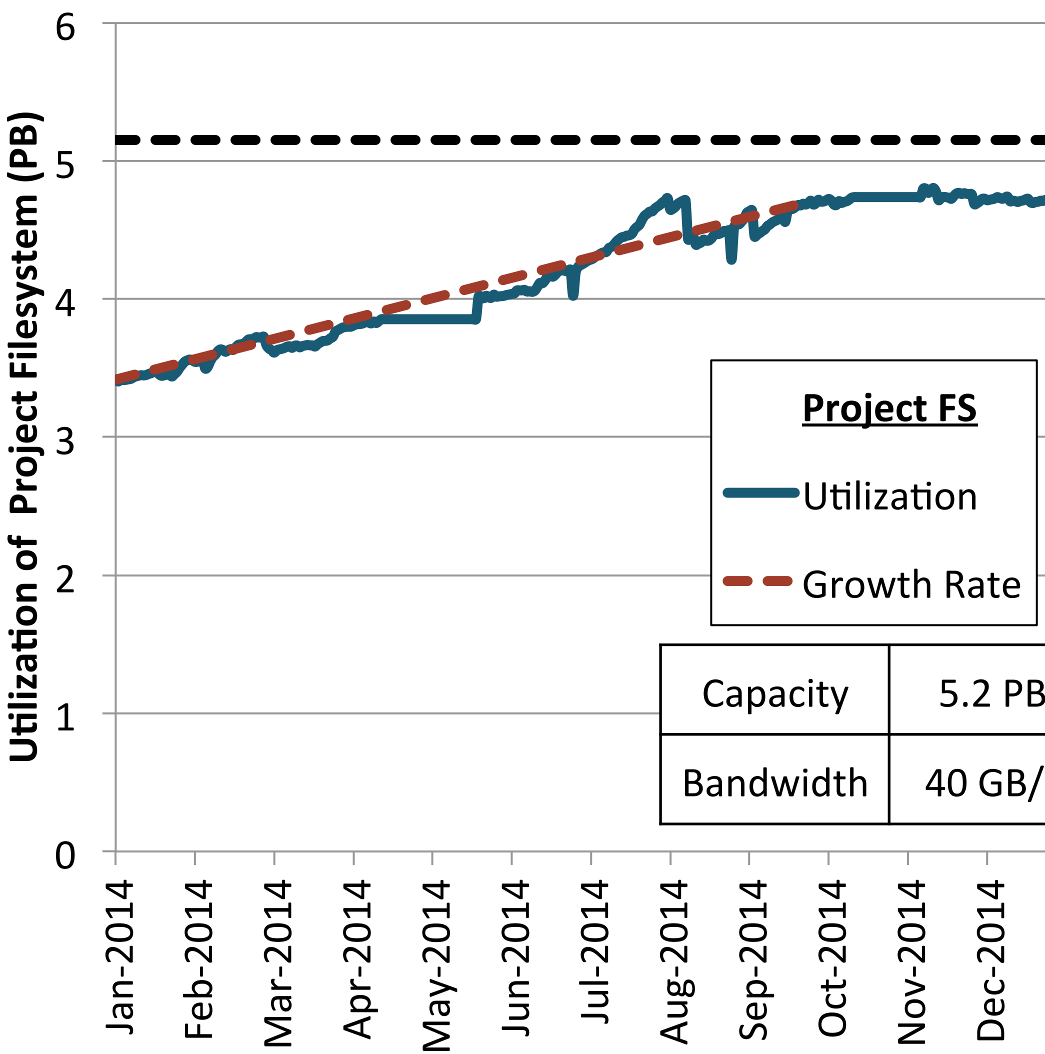
Filesystem	/scratch1	/scratch2	/scratch3	Total
Capacity (TB)	2100	2100	3200	7400
Bandwidth (GB/s)	48	48	72	168

- **Linear growth of /scratch1 and /scratch2**
  - 12 week purge policy
  - 1 TB quota per user
- **/scratch3 growth is less predictable.**
  - Piecewise linear?
  - 8 week purge policy
  - No quota
  - Fills more than 2x faster than /scratch1 or /scratch2
- **96% of data written to scratch is for temporary use.**
  - Average write volume is ~300 TB/day.
  - Aggregate *growth* of data stored is ~10 TB per day.

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

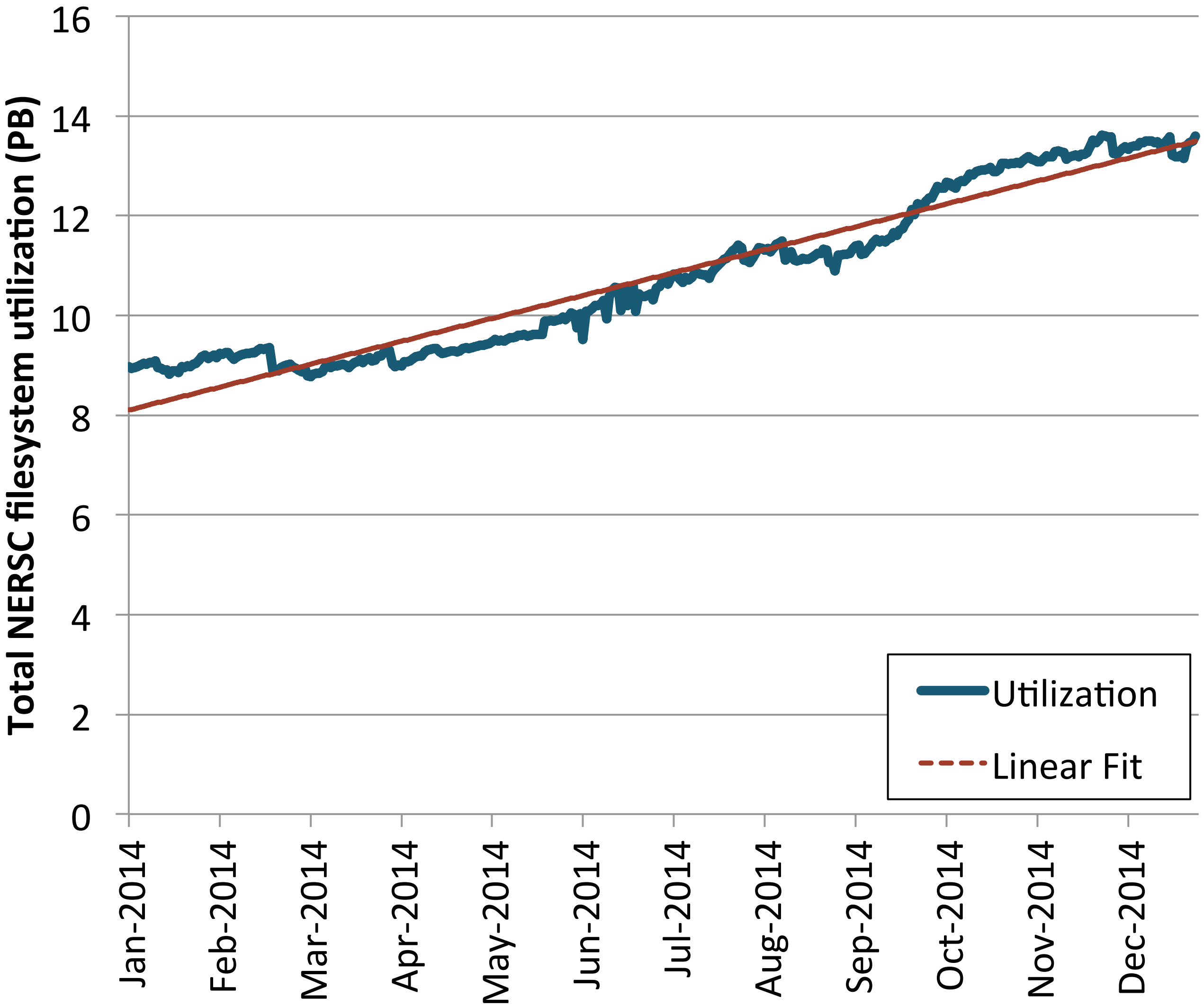
Project filesystem utilization increases 5 TB/day.



Capacity	5.2 PB
Bandwidth	40 GB/s

- **"Project" is a large, permanent, medium performance filesystem.**
- **Project directories are intended to facilitate sharing data among users and across NERSC systems.**
- **Linear growth**
  - No purge policy
  - 1 TB quota per project

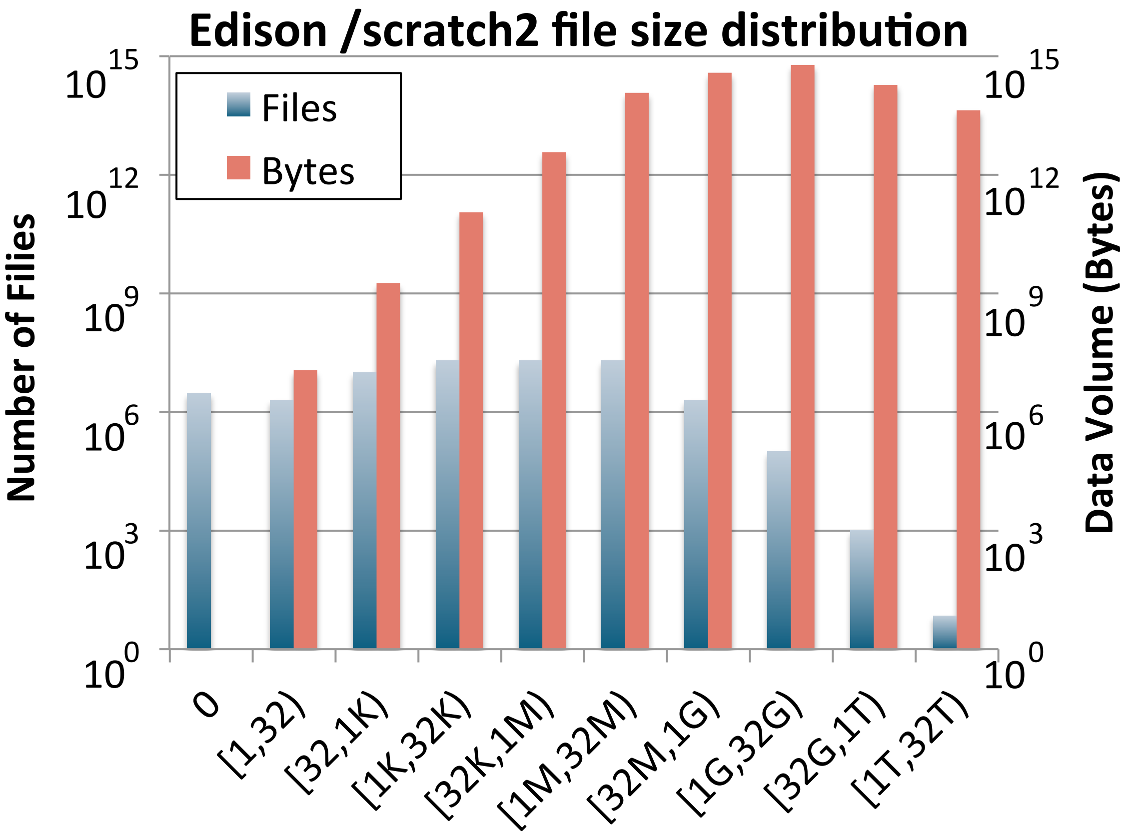
Total NERSC file-system utilization increases 15 TB/day.



- **Linear growth**
  - Summed over filesystems
  - Various quota and purge policies

	Capacity (TB)
Global homes	246
Global project	5150
Global projectb	2620
Global scratch	3600
Hopper scratch	1117
Hopper scratch2	1106
Edison scratch1	2100
Edison scratch2	2100
Edison scratch3	3200

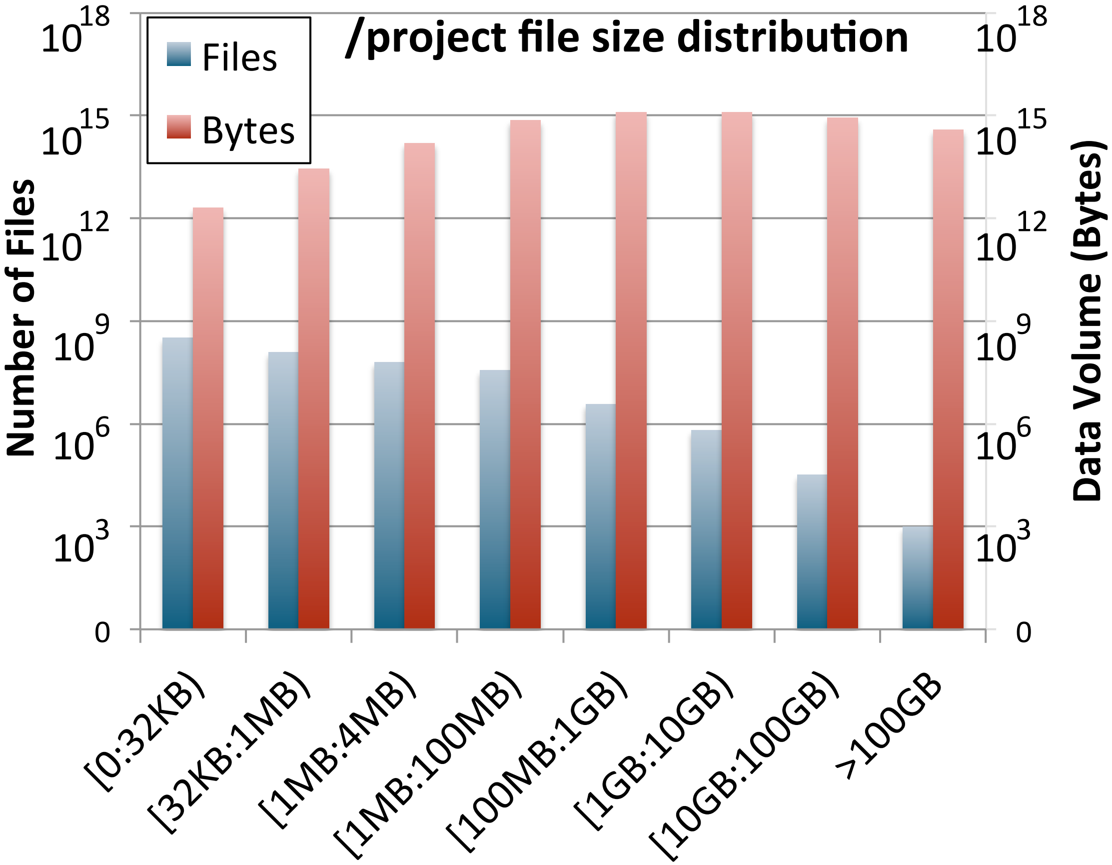
Files on Edison's scratch filesystems are generally small.



- **Average size: 9.4 MB**
- **Most (70%) files smaller than the 1 MB Lustre stripe size.**
- **Vast majority (>97%) of files smaller than 32 MB.**
- **Most (>90%) data is in files larger than 1MB.**

Total Count	Total Volume	Min	Max
91 M	821 TB	0 B	5 TB

File sizes on /project are similar to Edison's /scratch2.



- **Average size: 8.1 MB**
- **Most (80%) files smaller than the 1 MB.**
- **Most (>90%) data is in files larger than 1MB.**

Total Count	Total Volume	Min	Max
553M	4278TB	0 B	>1 TB

Storage and I/O summary

- I/O metadata and bandwidth performance are highly variable.
- Users seldom see the I/O rates they expect.
- Edison's maximum daily write volume is about twice its memory capacity. Hopper reads more data than Edison, sometimes 3x memory capacity per day.
- About 70% of the workload seems to use checkpoint/restart to cope with queue walltime limits.
- Filesystem utilization increases roughly linearly (15 TB/day).
- Most files (70%) are smaller than 1 MB. Most data (>90%) is in files larger than 1 MB.

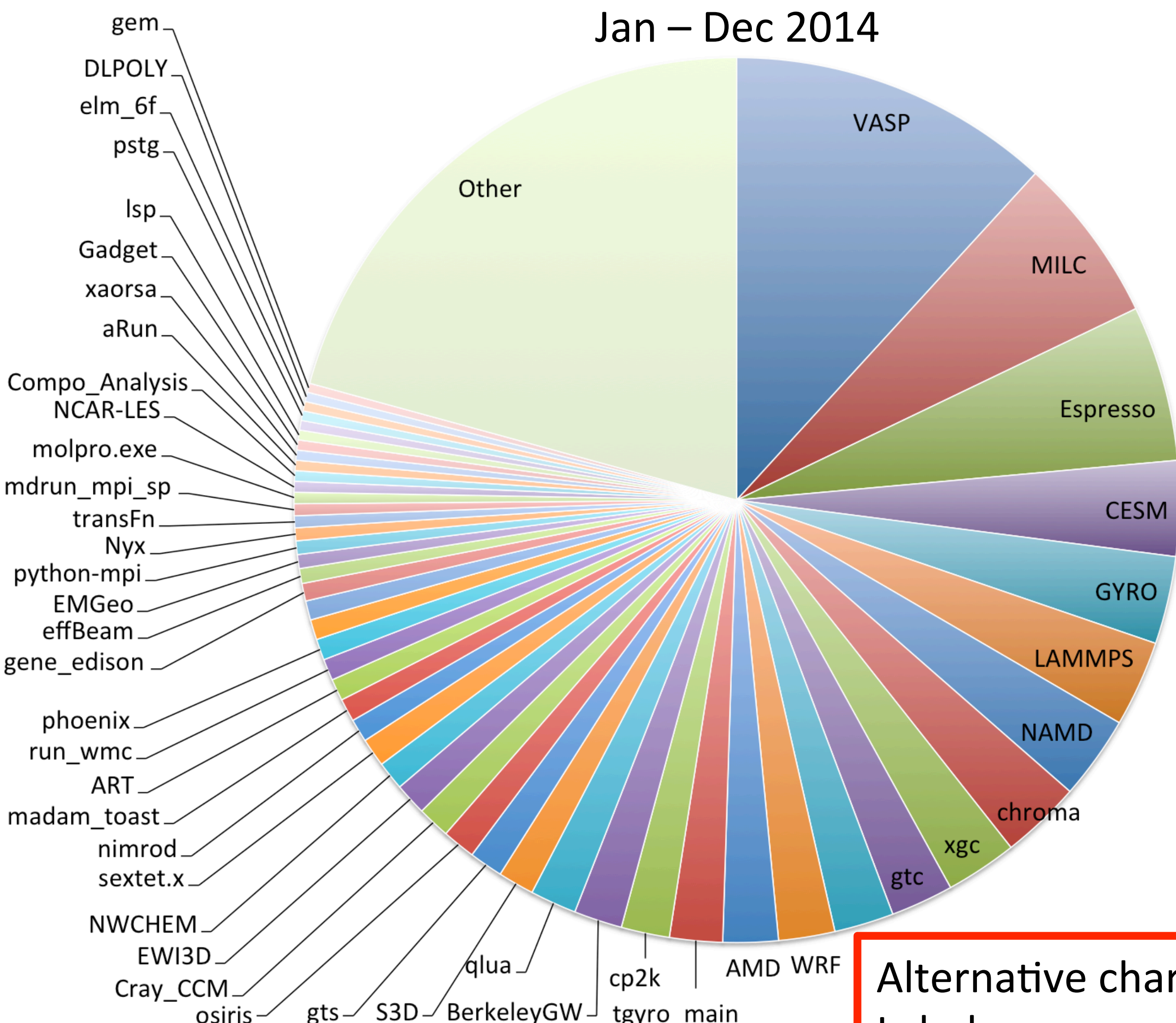
Conclusions

- **NERSC supports many users, domains and algorithms, and has a broad scientific impact.**
- **Most codes are still written Fortran, C++, or C, with MPI parallelism. OpenMP thread usage is 20%.**
  - For large jobs, any OpenMP inefficiencies are outweighed by MPI scalability issues.
  - Among threaded codes, the dominant thread concurrency matches the NUMA domain size.
- **Few Edison users are constrained by memory capacity.**
  - Half of the Edison workload will run in Cori's 16 GB HBM without modification.
- **Users seldom achieve large fractions of I/O bandwidth on scratch filesystems.**
  - Checkpoint – restart is common.
  - Maximum daily write volume is about 2x memory capacity.
  - Filesystem utilization grows steadily at 15 TB/day.

National Energy Research Scientific Computing Center

Over 650 applications run on NERSC resources.

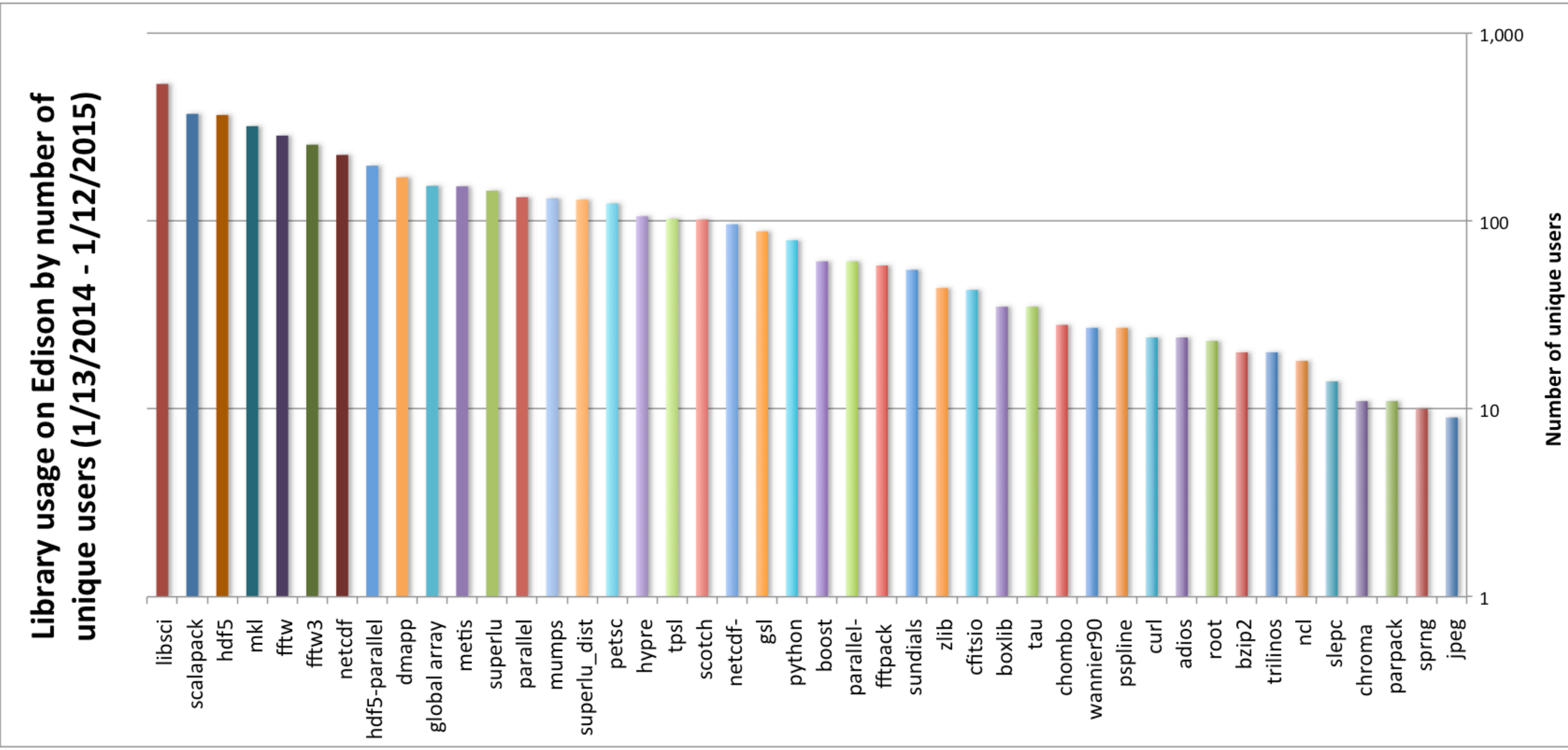
Top Application codes on Hopper and Edison by hours used.
Jan – Dec 2014

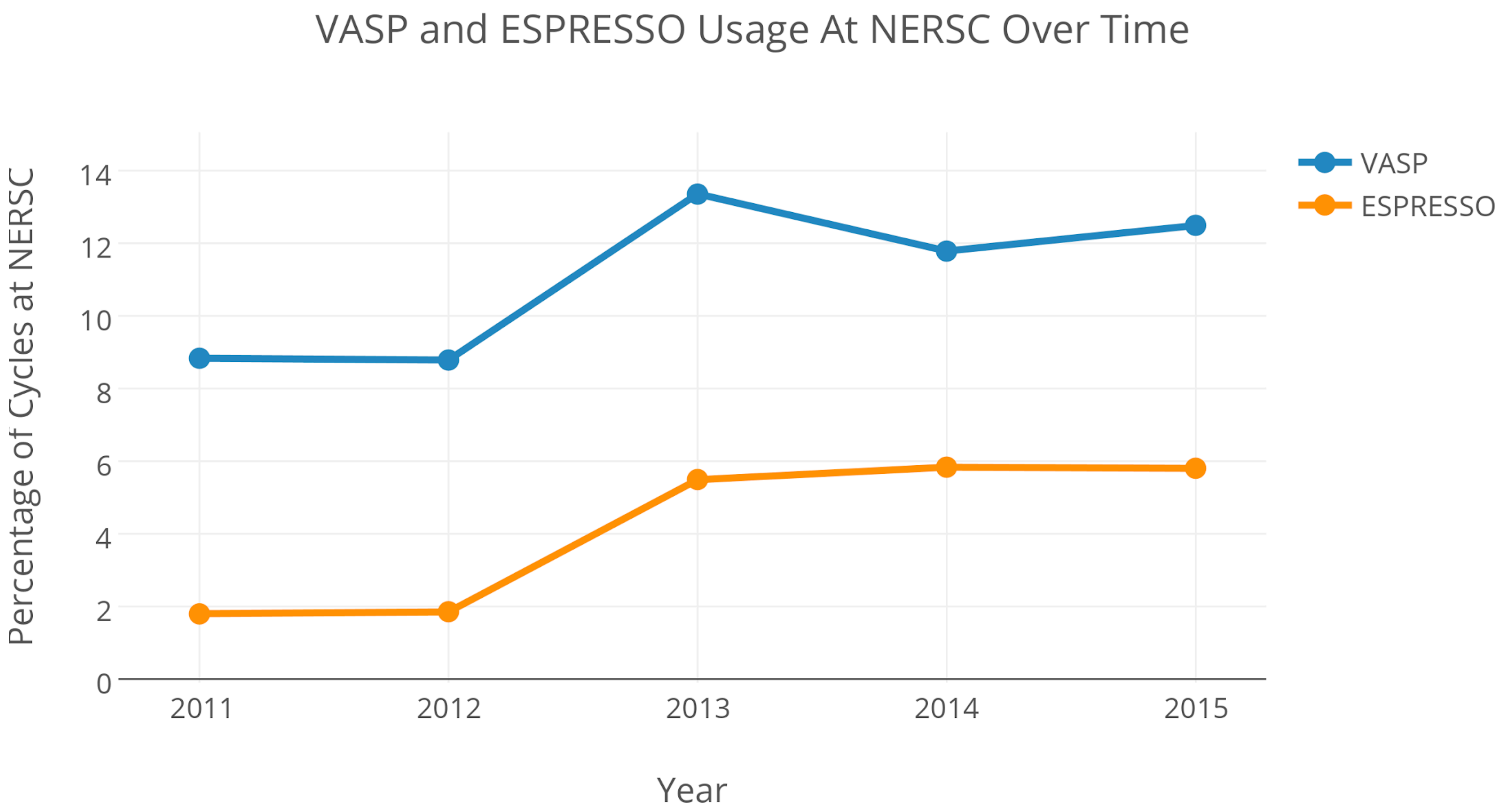


- **10 codes make up 45% of workload**
- **25 codes make up 66% of workload**
- **50 codes make up 80% of workload**
- **Remaining codes (over 600) make up 20% of workload.**

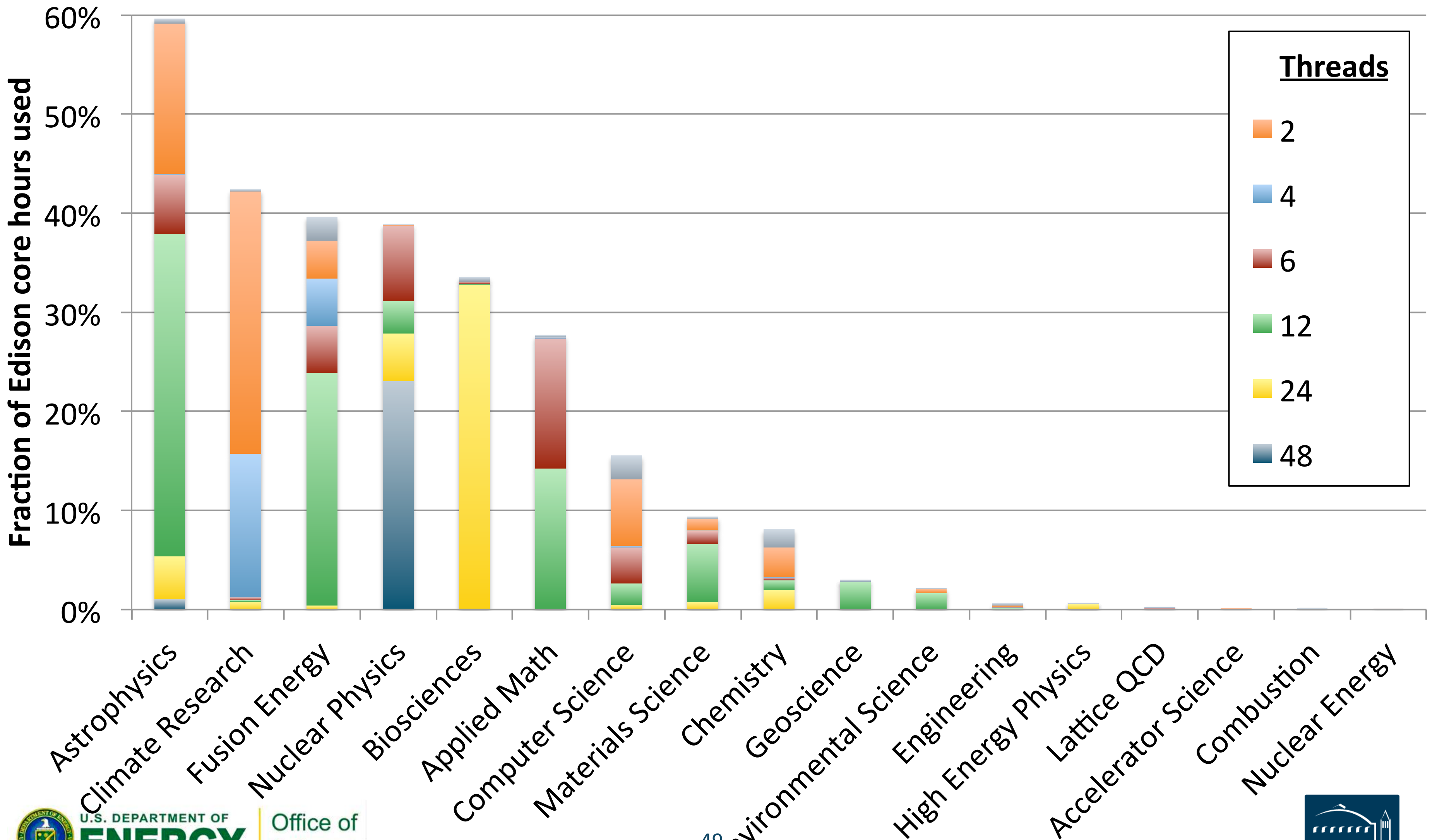
Alternative chart format;
Labels are more readable & assignable,
but pie size does not match format of other slides.

NERSC's broad workload relies on optimized libraries to maximize performance.

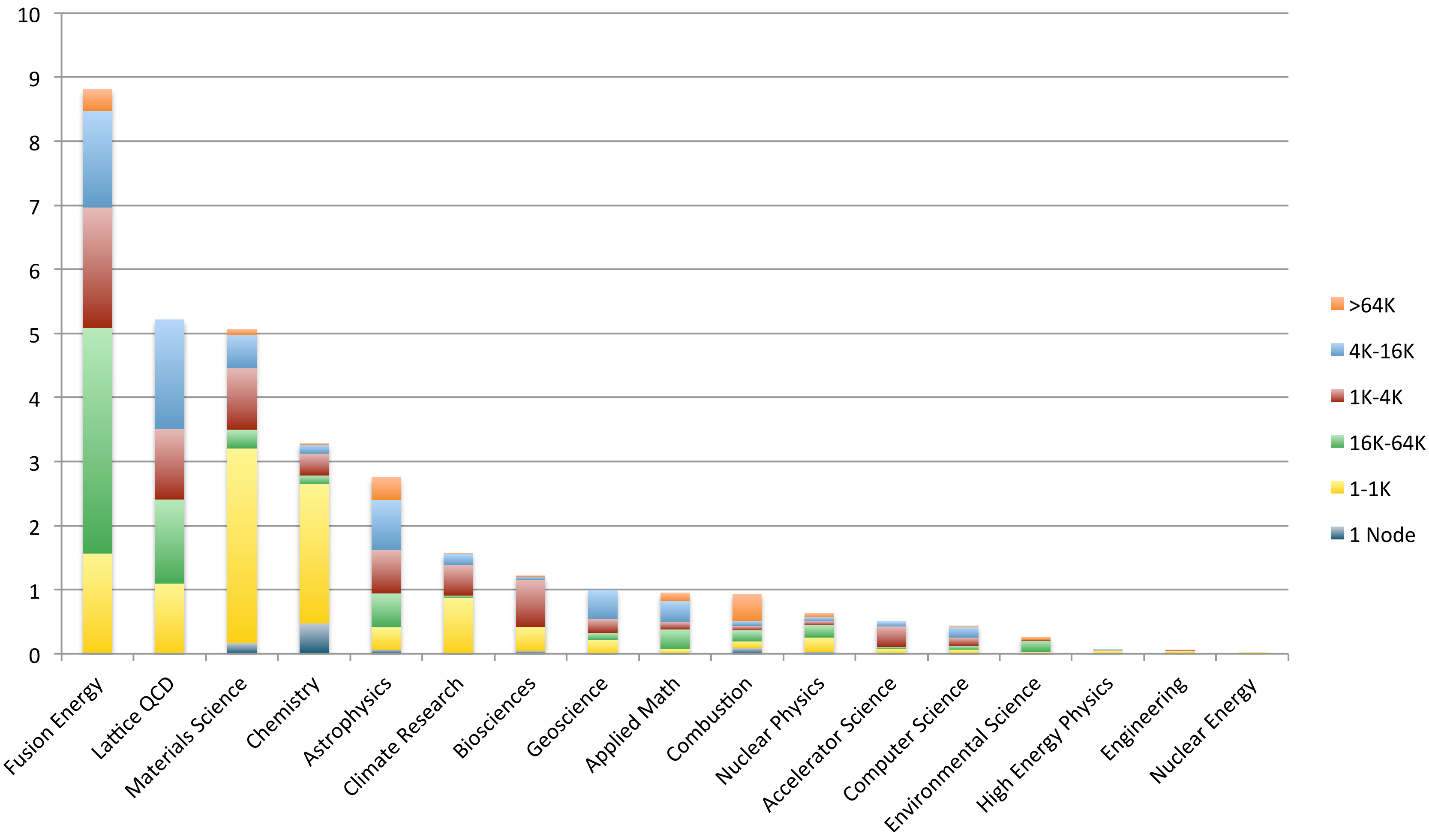



VASP and ESPRESSO Usage At NERSC Over Time
Percentage of Cycles at NERSC
0
2
4
6
8
10
12
14
2011
2012
2013
2014
2015
Year
VASP
ESPRESSO

Adoption of threads varies across disciplines.

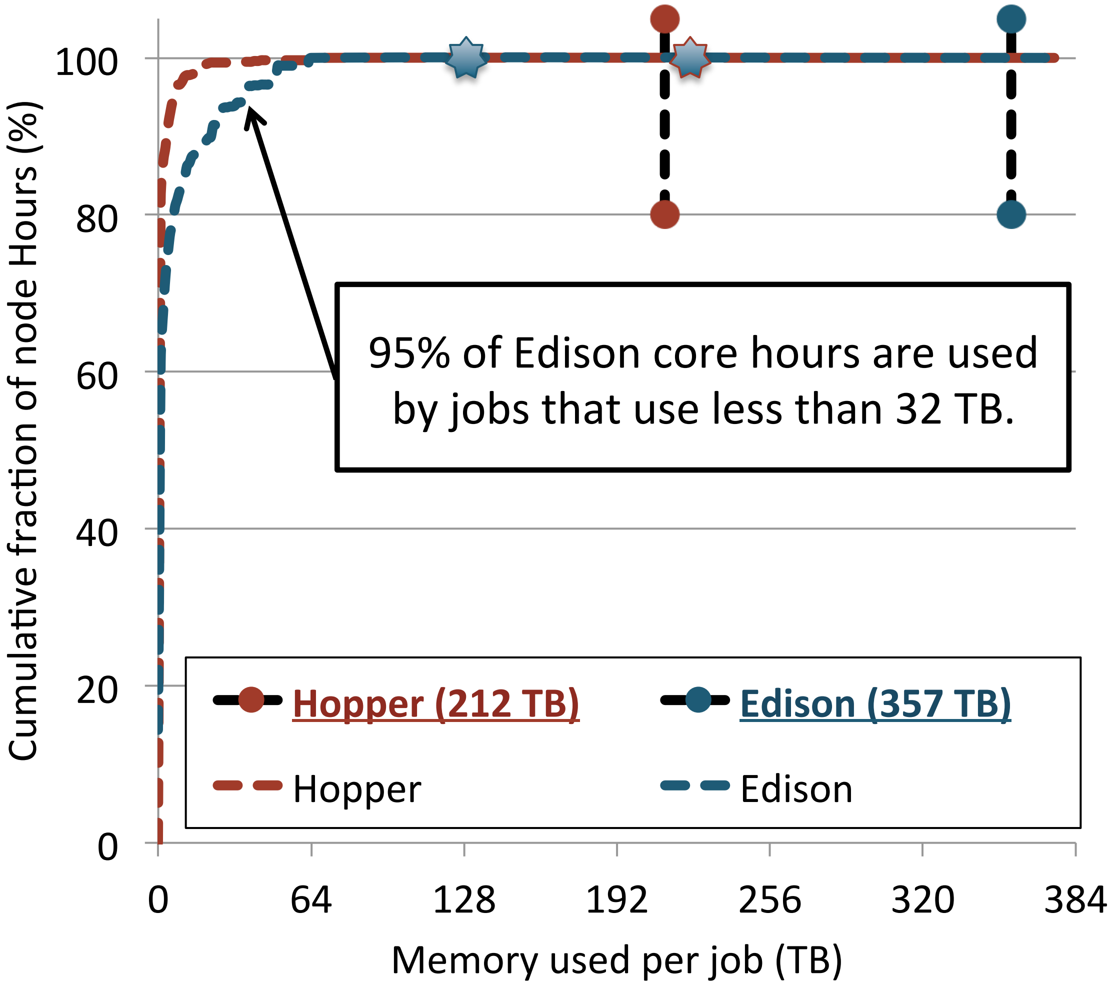


Science domains have different concurrency needs.



U.S. DEPARTMENT OF ENERGY | Office of Science

Users choose Edison for running jobs with large aggregate memory footprints.

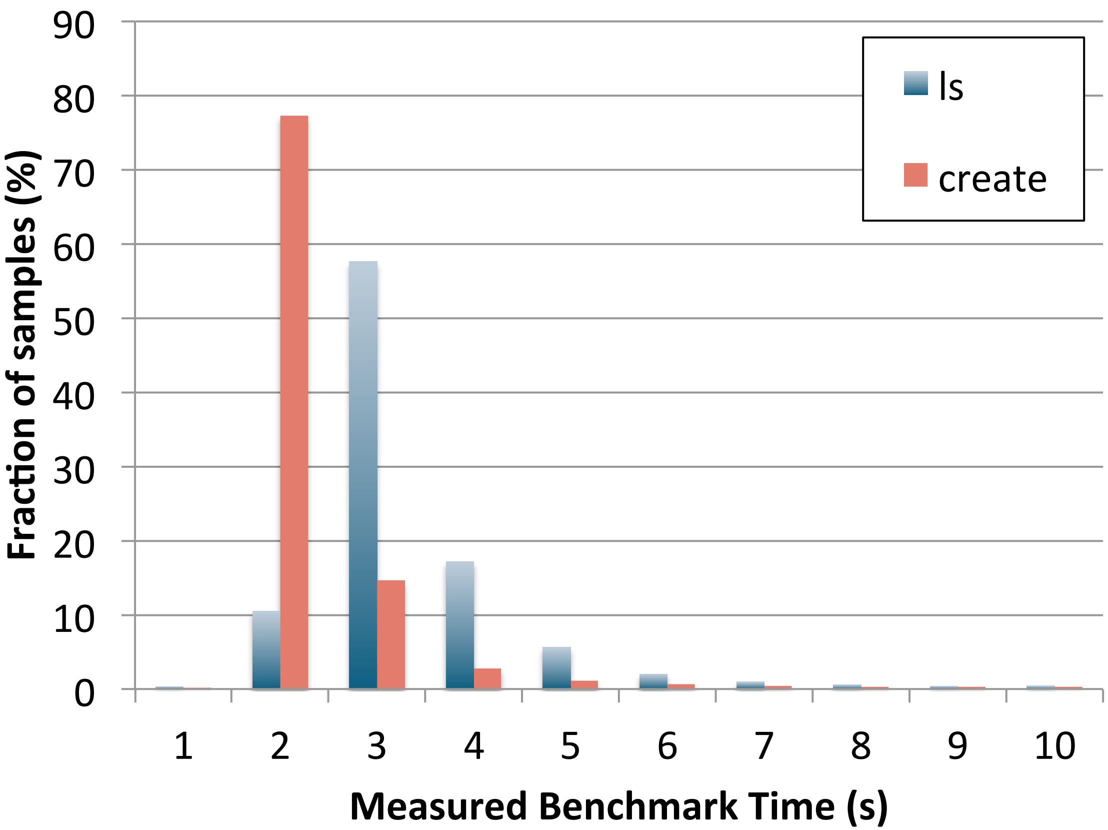


- **When given more powerful nodes and networks, users take advantage of increased memory (but not always at full-system scale).**
- **Memory capacity does not constrain Edison's largest jobs.**
  - Largest job uses only 2/3 memory; 10th largest uses 1/3 memory.
  - Edison's largest jobs could not fit on Hopper.

U.S. DEPARTMENT OF ENERGY | Office of Science

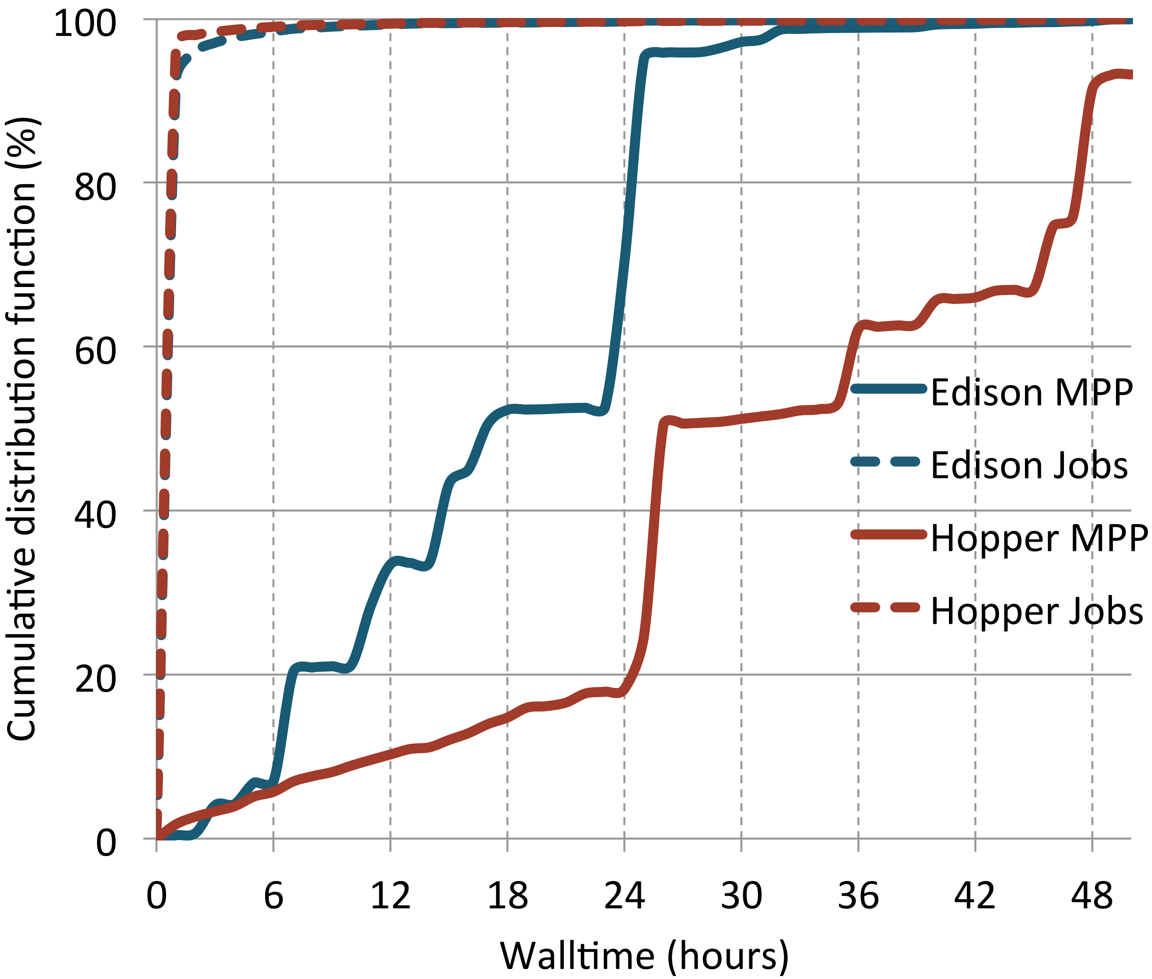
BERKELEY LAB

More reliable metadata performance would improve application performance variation.



- Cron job times "ls" and file creation every five minutes to test I/O metadata performance on Edison's scratch1 filesystem.
- Benchmarks normally complete in 2 or 3 seconds.
- More than one in five tests are significantly slower.
- Both benchmarks have long tails stretching to 300s.

Much of the NERSC workload relies on checkpoint-restart functionality.



- **A large fraction (70%) of core hours is consumed by jobs that reach the wallclock limit.**
  - Steps in plot correspond to queue limits.
  - This is only 0.5% of jobs.
- **Users want longer queues**

  (and shorter wait times)
- **95% of jobs run for less than one hour.**